Review of Economic Studies (2018) **85**, 1029–1069 doi:10.1093/restud/rdx033

Advance access publication 5 June 2017

What Motivates Effort? Evidence and Expert Forecasts

STEFANO DELLAVIGNA
UC Berkeley and NBER

and

DEVIN POPE
U Chicago and NBER

First version received August 2016*; Editorial decision March* 2017*; Accepted May* 2017 (*Eds.*)

How much do different monetary and non-monetary motivators induce costly effort? Does the effectiveness line up with the expectations of researchers and with results in the literature? We conduct a large-scale real-effort experiment with eighteen treatment arms. We examine the effect of (1) standard incentives; (2) behavioural factors like social preferences and reference dependence; and (3) non-monetary inducements from psychology. We find that (1) monetary incentives work largely as expected, including a very low piece rate treatment which does not crowd out effort; (2) the evidence is partly consistent with standard behavioural models, including warm glow, though we do not find evidence of probability weighting; (3) the psychological motivators are effective, but less so than incentives. We then compare the results to forecasts by 208 academic experts. On average, the experts anticipate several key features, like the effectiveness of psychological motivators. A sizeable share of experts, however, expects crowd-out, probability weighting, and pure altruism, counterfactually. As a further comparison, we present a meta-analysis of similar treatments in the literature. Overall, predictions based on the literature are correlated with, but underperform, the expert forecasts.

Key words: Behavioural Economics, Experts, Real effort, Crowd-out, Charity, Probability weighting.

JEL Codes: D03, C93

1. INTRODUCTION

Monetary incentives have long been used as a way to change behaviour. More recently, policy-makers, researchers, and businesses have turned to behavioural economics and psychology for additional levers, for example with the formation of Behavioural Science Units.

A criticism of this approach is that there are too many potential levers to change behaviour, without a clear indication of their relative effectiveness. Different dependent variables and dissimilar participant samples make direct comparisons of effect sizes across various studies difficult. Given the disparate evidence, it is not clear whether even behavioural experts can determine the relative effectiveness of various interventions in a particular setting.

In this article, we run a large pre-registered experiment that allows us to compare the effectiveness of multiple treatments within one setting. We focus on a real-effort task with treatments including monetary incentives and non-monetary behavioural motivators. The

treatments are, as much as possible, model-based, so as to relate the findings to behavioural models and estimate the behavioural parameters.

In addition to providing evidence on the efficacy of various treatments, we also elicit forecasts from academic experts on the effectiveness of the treatments. We thus capture the beliefs of the research community on various behavioural topics. The forecasts also allow us to measure in which direction, and how decisively, the results diverge from such beliefs.

Turning to the details, we recruit 9,861 participants on Amazon Mechanical Turk (MTurk)—an online platform that allows researchers to post small tasks that require a human to perform. MTurk has become very popular for experimental research in marketing and psychology (Paolacci and Chandler, 2014) and is increasingly used in economics as well (*e.g.* Kuziemko *et al*., 2015).[1] The limited cost per subject and large available population on MTurk allow us to run eighteen treatments with over 500 subjects in each treatment arm.

The task for the subjects is to alternately press the "a" and "b" buttons on their keyboards as quickly as possible for ten minutes. The eighteen treatments attempt to motivate participant effort using (1) standard incentives, (2) non-monetary psychological inducements, and (3) behavioural factors such as social preferences, present bias, and reference dependence.

We present three main findings about performance. First, monetary incentives have a strong and monotonic motivating effect: compared to a treatment with no piece rate, performance is 33% higher with a 1-cent piece rate, and another 7% higher with a 10-cent piece rate. A simple model of costly effort estimated on these three benchmark treatments predicts performance very well not only in a fourth treatment with an intermediate (4-cent) piece rate, but also in a treatment with a very low (0.1-cent) piece rate that could be expected to crowd out motivation. Instead, effort in this very-low-pay treatment is 24% higher than with no piece rate, in line with the predictions of a model of effort for this size of incentive.

Second, non-monetary psychological inducements are moderately effective in motivating the workers. The three treatments increase effort compared to the no-pay benchmark by 15–21%, a sizeable improvement especially given that it is achieved at no additional monetary cost. At the same time, these treatments are less effective than any of the treatments with monetary incentives, including the one with very low pay. Among the three interventions, two modelled on the social comparison literature and one on task significance (Grant, 2008), a Cialdini-type comparison (Schultz *et al*., 2007) is the most effective.

Third, the results in the behavioural treatments are partly consistent with behavioural models of social preferences, time preferences, and reference dependence, with important nuances. Treatments with a charitable giving component motivate workers, but the effect is independent of the return to the charity (1-cent or 10-cent piece rate). We also find some, though quantitatively small, evidence of a reciprocal gift-exchange response to a monetary "gift".

Turning to time preferences, treatments with payments delayed by 2 or 4 weeks induce less effort than treatments with immediate pay, for a given piece rate, as expected. However, the decay in effort is exponential, not hyperbolic, in the delay, although the confidence intervals of the estimates do not rule out significant present bias.

We also provide evidence on two key components of reference dependence, loss aversion and overweighting of small probabilities. Using a claw-back design (Hossain and List, 2012), we find a larger response to an incentive framed as a loss than as a gain, though the difference is not significant. Probabilistic incentives as in Loewenstein *et al*. (2007), though, induce less effort than a deterministic incentive with the same expected value. This result is not consistent

1. A legitimate question is the comparability of studies run on MTurk versus in more standard laboratory or field settings. Evidence suggests that MTurk findings are generally qualitatively and quantitatively similar (Horton *et al*., 2011) to findings in more traditional platform.

with overweighting of small probabilities (assuming the value function is linear or moderately concave).

In the second stage of this project, we measure the beliefs of academic experts about the effectiveness of the treatments. We surveyed researchers in behavioural economics, experimental economics, and psychology, as well as some non-behavioural economists. We provided the experts with the results of the three benchmark treatments with piece-rate variation to help them calibrate how responsive participant effort was to different levels of motivation in this task. We then ask them to forecast the effort participants exerted in the other fifteen treatment conditions. To ensure transparency, we pre-registered the experiment and we ourselves did not observe the results of the fifteen treatment conditions until after the collection of expert forecasts.

Out of 312 experts contacted, 208 experts provided a complete set of forecasts. The broad selection and the 66% rate ensure a good coverage of behavioural experts.

The experts anticipate several results, and in particular the effectiveness of the psychological inducements. Strikingly, the average forecast ranks in the exact order the six treatments without private performance incentives: two social comparison treatments, a task significance treatment, the gift exchange treatment, and two charitable giving treatments.

At the same time, the experts mispredict certain features. The largest deviation between the average expert forecast and the actual result is for the very-low-pay treatment, where experts on average anticipate a 12% crowd out, while the evidence indicates no crowd out. In addition, while the experts predict correctly the average effort in the charitable giving treatments, they expect higher effort when the charity earns a higher return; the effort is instead essentially identical in the two charitable treatments. The experts also overestimate the effectiveness of the gift exchange treatment by 7%.

Regarding the other behavioural treatments, in the delayed-payout treatments the experts predict a pattern of effort consistent with present bias, while the evidence is most consistent with exponential discounting. The experts expect the loss framing to have about the same effect as a gain framing with twice the incentives, consistent with the Tversky and Kahneman (1991) calibration and largely in line with the MTurker effort. The experts also correctly expect the probabilistic piece rates to underperform the deterministic piece rate with same expected value, though they still overestimate the effectiveness of the probabilistic incentives.

How do we interpret the differences between the experimental results and the expert forecasts? We consider three classes of explanations: *biased literature*, *biased context*, and *biased experts*. In the first explanation, *biased literature*, the published literature upon which the experts rely is biased, perhaps due to its sparsity or some form of publication bias. In the second explanation, *biased context*, the literature itself is not biased, but our experimental results are unusual and differ from the literature due to our particular task or subject pool. In the third explanation, *biased experts*, the forecasts are in error because the experts themselves are biased—perhaps due to the experts failing to rely on or not knowing the literature.

With these explanations in mind, we present a meta-analysis of papers in the literature.[2] We include lab and field experiments on effort (broadly construed) that include treatment arms similar to ours. The resulting data set includes 42 papers covering 8 of the 15 treatment comparisons.[3] For each treatment comparison, we compute the weighted average effect in standard deviation units (Cohen's d) from the literature.

We stress three features of this data set. First, we found only one paper that uses MTurk subjects for a similar treatment; thus, the experts could not rely on experiments with a comparable sample. Second, nearly all papers contain only one type of treatment; papers such as ours and

2. This meta-analysis was not part of the pre-analysis plan. We are grateful to the referees for the suggestion.
3. Some treatments are not included because we could not identify relevant papers for the meta-analysis.

Bertrand *et al*. (2010) comparing a number of behavioural interventions are uncommon. Third, for most treatments we found only a few papers, sometimes little-known studies outside economics, including for classical topics such as probability weighting.[4] Thus, an expert who wanted to consult the literature could not simply look up one or two familiar papers.

We find evidence consistent with all three classes of explanations. In the very-low-pay condition, both the experts and the literature substantially underpredict the effort. This could be a result of a *biased literature* or a *biased context* (and experts are unable to adapt the results from the literature to our unique context). In another example, the literature-based forecasts accurately predict that the low-return and the high-return charity treatments will induce similar effort, whereas the experts predict higher effort levels when the return to charity increases. This treatment provides evidence in favour of a *biased expert* account.

In general, our simple meta-analysis proves to be a worse predictor of the results than the experts: the average absolute deviation between predictions and results is more than twice as large for the literature-based predictions than for the expert forecasts. This difference gets even larger if the meta-analysis weighs papers based on their citation count. This helps put in perspective the remarkable forecasting accuracy of the experts.

In the final part of the article, we exploit the model-based design to estimate the behavioural parameters underlying the observed MTurk effort and the expert forecasts. With respect to social preferences, the effort supports a simple "warm glow" model, while the median expert expects a pure altruism model. Regarding the time preferences, the median expert expects a β of 0.76, in line with estimates in the literature, while the point estimate for β from the MTurker effort (while noisy) is around 1. On reference dependence, assuming a value function calibrated as in Tversky and Kahneman (1991), we find *under*weighting of small probabilities, while the median expert expects (modest) overweighting. If we jointly estimate the curvature as well, the data can accommodate probability weighting, but for unrealistic values of curvature. Finally, we back out the loss aversion parameter using a linear approximation.

We explore complementary findings on expert forecasts in a companion paper (DellaVigna and Pope, forthcoming). We present measures of expert accuracy, comparing individual forecasts with the average forecast. We also consider determinants of accuracy and compare the predictions of academic experts to those of other groups: PhDs, undergraduates, MBAs, and MTurkers. We also examine beliefs of experts about their own expertise and the expertise of others. Thus, the companion paper focuses on what makes a good forecaster, while this article is focused on behavioural motivators and the beliefs that experts hold about the behavioural treatments.

Our findings relate to a vast literature on behavioural motivators.[5] Several of our treatments have parallels in the literature, such as Imas (2014) and Tonin and Vlassopoulos (2015) on effort and charitable giving. Two main features set our study apart. First, we consider the behavioural motivators in a common environment, allowing us to measure the relative effectiveness. Second, we compare the effectiveness of behavioural interventions with the expert expectations.

The emphasis on expert forecasts ties this article to a small literature on forecasts of research results.[6] Coffman and Niehaus (2014) survey seven experts on persuasion, while Sanders *et al*.

4. There is a large experimental literature on probability weighting, but on lottery choices, not on effort tasks.

5. Among other papers, our treatments relate to the literature on pro-social motivation (Andreoni, 1989 and 1990), crowd-out (Gneezy and Rustichini, 2000), present-bias (Laibson, 1997; O'Donoghue and Rabin, 1999), and reference dependence (Kahneman and Tversky, 1979; Koszegi and Rabin, 2006).

6. There is a larger literature on forecasting about topics other than research results, *e.g*. the Good Judgment Project on national security (Tetlock and Gardner, 2015; Mellers *et al*., 2015). Several surveys, like the IGM Economic Expert panel, elicit opinions of experts about economic variables, such as inflation or stock returns.

(2015) ask twenty-five faculty and students from two universities questions on fifteen select experiments run by the UK Nudge Unit. Groh *et al*. (2015) elicit forecasts on an RCT from audiences of four academic presentations. Erev *et al*. (2010) ran a competition among laboratory experimenters to forecast the result of a laboratory experiment using learning models trained on data. These complementary efforts suggest the need for a more systematic collection of expert beliefs about research findings.

We are also related to a recent literature on transparency in the social sciences (*e.g*. Simmons *et al*., 2011; Vivalt, 2016; Banerjee *et al*., 2016), including the use of prediction markets[7] to capture beliefs about the replicability of experimental findings (Dreber *et al*., 2015; Camerer *et al*., 2016). We emphasize the complementarity, as our study examines a novel real-effort experiment building on behavioural models, while the Science Prediction Market concerns the exact replication of existing protocols.

Our article also adds to a literature on *structural behavioural economics*.[8] A unique feature is that we compare estimates of behavioural parameters in the data to the beliefs of experts.

The article proceeds as follows. In Section 2 we motivate the treatments in light of a simple costly-effort model, and in Section 3 we present the design. We present the treatment results in Section 4, the evidence on forecasts in Section 5, and the meta-analysis in Section 6. In Section 7 we derive the implied behavioural parameters and in Section 8 we conclude.

2. TREATMENTS AND MODEL

In this section, we motivate the eighteen treatments in the experiment (Table 1) in light of a simple model of worker effort. As we will describe in more detail in Section 3, the MTurk workers have 10 minutes to complete a real-effort task (pressing a-b keys), with differences across the treatments in incentives and behavioural motivators. The model of costly effort, which we used to design the experiment and is registered in the pre-analysis plan, ties the eighteen treatments to key behavioural models, like present bias and reference dependence.

2.1. *Piece rates*

The first four treatments involve variation in the piece rate received by experiment participants to push buttons. (The piece rate is in addition to the advertised compensation of a $1 flat fee for completing the task.) In the first treatment subjects are paid no piece rate ("*Your score will not affect your payment in any way*"). In the next three treatments there is a piece rate at 1 cent ("*As a bonus, you will be paid an extra 1 cent for every 100 points that you score*"), 10 cents ("*As a bonus, you will be paid an extra 10 cents for every 100 points that you score*"), and 4 cents ("*As a bonus, you will be paid an extra 4 cents for every 100 points that you score*"). The 1-cent piece rate per 100 points is equivalent to an average extra 15–25 cents, which is a sizeable pay increase for a 10-minute task in MTurk. The 4-cent piece rate and, especially, the 10-cent piece rate represent substantial payment increases by MTurk standards. These stated piece rates are the only differences across the treatments.

7. See, for example, Snowberg *et al*. (2007) on prediction markets.

8. Papers include Laibson *et al*. (2007), Conlin *et al*. (2007), DellaVigna *et al*. (2012), Barseghyan *et al*. (2013), DellaVigna *et al*. (2015).

TABLE 1
Summary of eighteen treatments

Category (1)	Treatment Wording (2)	Parameter (3)	Cites (4)
Piece rate	"Your score will not affect your payment in any way."		
	As a bonus, you will be paid an extra **1 cent** for every **100 points** that you score."		
	"As a bonus, you will be paid an extra **10 cents** for every **100 points** that you score."		
	"As a bonus, you will be paid an extra **4 cents** for every **100 points** that you score."		
Pay enough or don't pay	"As a bonus, you will be paid an extra **1 cent** for every **1,000 points** that you score."	ΔS_{CO} (crowd out)	Deci, 1971; Gneezy and Rustichini, 2000
Social preferences: charity	"As a bonus, the Red Cross charitable fund will be given **1 cent** for every 100 points that you score."	α (altruism) and a (warm glow)	Andreoni, 1989 and 1990; Becker, 1974; Imas, 2014
	"As a bonus, the Red Cross charitable fund will be given **10 cents** for every 100 points that you score."		
Social preferences: gift exchange	"In appreciation to you for performing this task, you will be paid a **bonus** of **40 cents**. Your score will not affect your payment in any way."	ΔS_{GE}	Fehr *et al*., 1993; Gneezy and List, 2006
Discounting	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **two weeks** from today."	β, δ (impatience parameters)	Laibson, 1997; O'Donoghue and Rabin, 1999; Andreoni and Sprenger, 2012; Augenblick *et al*., 2015
	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **four weeks** from today."		
Gains versus losses	"As a bonus, you will be paid an **extra 40 cents** if you score at least 2,000 points."	λ (loss aversion)	Kahneman and Tversky, 1979; Hossain and List, 2012; Fryer *et al*., 2012
	"As a bonus, you will be paid an **extra 40 cents**. However, you will **lose this bonus** (it will not be placed in your account) unless you score at least 2,000 points."		
	"As a bonus, you will be paid an **extra 80 cents** if you score at least 2,000 points."		
Risk aversion and probability weighting	"As a bonus, you will have a **1% chance** of being paid an **extra $1** for every 100 points that you score. One out of every 100 participants who perform this task will be randomly chosen to be paid this reward."	π(P) (probability weighting)	Kahneman and Tversky, 1979; Prelec, 1998; Wu and Gonzalez, 1996; Loewenstein *et al*., 2007
	"As a bonus, you will have a **50% chance** of being paid an **extra 2 cents** for every 100 points that you score. One out of two participants who perform this task will be randomly chosen to be paid this reward."		

(*continued*)

TABLE 1
Continued

Category (1)	Treatment Wording (2)	Parameter (3)	Cites (4)
Social comparisons	"Your score will not affect your payment in any way. In a previous version of this task, **many participants** were able to score more than **2,000 points.**"	ΔS_{SC}	Schultz *et al*., 2007
Ranking	"Your score will not affect your payment in any way. After you play, we will show you **how well you did relative** to other participants who have previously done this task."	ΔS_R	Maslow, 1943; Bandiera *et al*., 2013; Ashraf *et al*., 2014
Task significance	"Your score will not affect your payment in any way. We are interested in how fast people choose to press digits and we would like you to do your very best. So **please try as hard as you can**."	ΔS_{TS}	Grant, 2008

Notes: The Table lists the eighteen treatments in the Mturk experiment. The treatments differ just in one paragraph explaining the task and in the vizualization of the points earned. Column (2) reports the key part of the wording of the paragraph. For brevity, we omit from the description the sentence "This bonus will be paid to your account within 24 hours" which applies to all treatments with incentives other than in the Time Preference ones where the payment is delayed. Notice that the bolding is for the benefit of the reader of the Table and was not used in the treatment description on MTurk. Column (1) reports the conceptual grouping of the treaments, Column (3) reports the parameters in the model related to the treatment, and Column (4) reports some key references for the treatment.

The 0-cent, 1-cent, and 10-cent treatments provide evidence on the responsiveness of effort to incentives for this particular task. As such, we provide the results for these benchmark treatments to the experts so as to facilitate their forecasts of the other treatments. Later, we use the results for these treatments to estimate a simple model of costly effort and thus back out the behavioural parameters.

Formally, we assume that participants in the experiment maximize the return from effort e net of the cost of effort. Let e denote the number of points (*i.e*. alternating a-b presses). For each point e, the individual receives a piece-rate p as well as a non-monetary reward, $s>0$. The parameter s captures, in reduced form, a norm or sense of duty to put in effort for an employer, or gratitude for the \$1 flat payment for the 10-minute task. It could also capture intrinsic motivation or personal competitiveness from playing a game/puzzle like our task, or motivation to attain approval for the task.[9] This motivation is important because otherwise, for $s=0$, effort would equal zero in the no-piece rate treatment, counterfactually.

We assume a convex cost of effort function $c(e)$: $c'(e)>0$ and $c''(e)>0$ for all $e>0$. Assuming risk-neutrality, an individual solves

$$\max_{e\geq 0}(s+p)e-c(e), \tag{1}$$

leading to the solution (when interior) $e^*=c'^{-1}(s+p)$. Optimal effort e^* is increasing in the piece rate p and in the motivation s. We consider two special cases for the cost function, discussed further in DellaVigna *et al*. (2015). The first function, which we pre-registered, is the power cost function $c(e)=ke^{1+\gamma}/(1+\gamma)$, characterized by a constant elasticity of effort $1/\gamma$ with respect

9. While we granted approval for all effort levels, as promised, participants may have thought otherwise.

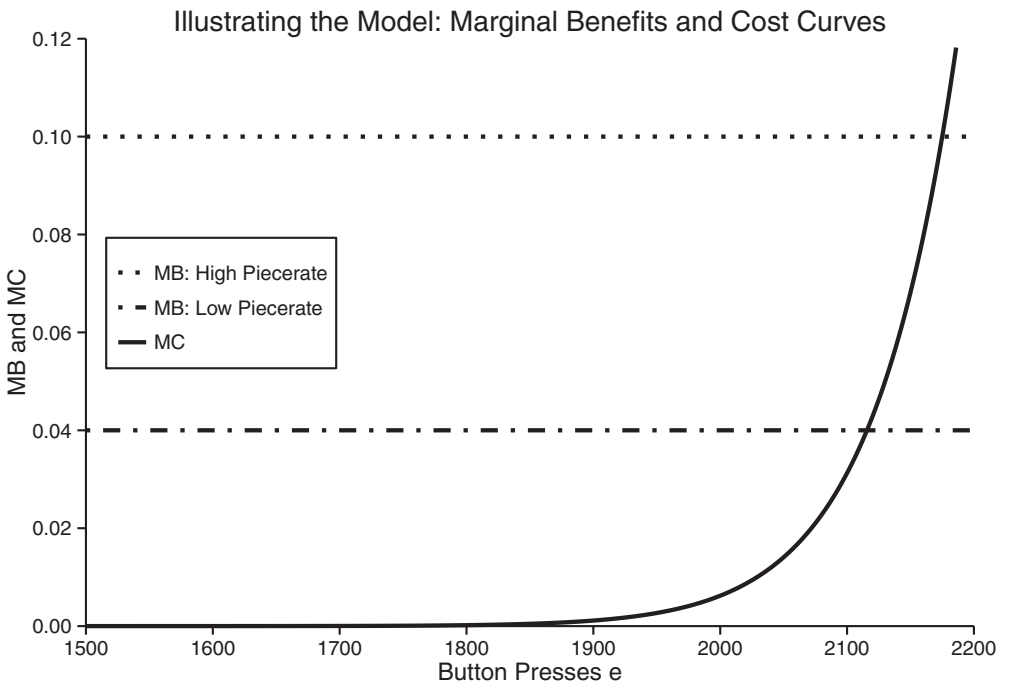

FIGURE 1

Model of effort determination, marginal benefit and marginal cost

Notes: Figure 1 plots the determination of the equilibrium effort at the intersection of marginal cost and marginal benefit. The different piece rate treatments shift the marginal benefit curve, holding the marginal cost curve constant.

to the value of effort. Under this assumption, we obtain

$$e^* = \left(\frac{s+p}{k}\right)^{1/\gamma}. \tag{2}$$

A plausible alternative is that the elasticity decreases as effort increases. A function with this feature is the exponential cost function, $C(e) = k\exp(\gamma e)/\gamma$, leading to solution

$$e^* = \frac{1}{\gamma}\log\left(\frac{s+p}{k}\right). \tag{3}$$

Under either function, the solution for effort has three unknowns, s, k, and γ which we can back out from the observed effort at different piece rates, as we do in Sections 4 and 7.

As Figure 1 illustrates, for a given marginal cost curve $c'(e)$ (black solid line), changes in piece rate p shift the marginal benefit curve $s+p$, plotted for two levels of piece rate p (dashed lines). The optimal effort $e^*(p)$ is at the intersection of the marginal cost and marginal benefit.

We stress two key simplifying assumptions. First, we assume that the workers are homogeneous, implying (counterfactually) that they would all make the same effort choice in a given treatment. Second, even though the piece rate is earned after a discrete number of points (100 points, or 1,000 points below), we assume that it is earned continuously so

as to apply the first-order conditions. We make these restrictive assumptions to ensure the model is simple enough to be estimated using just the three benchmark moments which the experts observe. In Section 7 we present an alternative estimation method which relaxes these assumptions.

2.2. *Very low pay*

Motivated by the crowd-out literature (Deci, 1971), we design a treatment with very low pay (Gneezy and Rustichini, 2000): "*As a bonus, you will be paid an extra 1 cent for every 1,000 points that you score.*" Even by MTurk standards, earning an extra cent upon spending several minutes on effortful presses is a very limited reward. Thus, it may be perceived as offensive and lead to lower effort. We model the treatment as corresponding to a piece rate $p = .001$, with a shift Δs_{CO} in motivation s:

$$e^*_{CO} = c'^{-1}(s + \Delta s_{CO} + p). \tag{4}$$

We should note that the task at hand is not necessarily an intrinsically rewarding task. As such, one may argue that the crowd-out literature does not predict reduced effort. Even under this interpretation, it is useful to compare the results to the expert expectations.

2.3. *Social preferences*

The next two treatments involve charitable giving: "*As a bonus, the Red Cross charitable fund will be given 1 cent for every 100 points that you score*" and "*as a bonus, the Red Cross charitable fund will be given 10 cents for every 100 points that you score.*" The rates correspond to the piece rates in the benchmark treatments, except that the recipient now is a charitable organization instead of the worker, similar to Imas (2014) and Tonin and Vlassopoulos (2015). The two treatments allow us to test (1) how participants feel about money for a charity versus money for themselves and (2) whether they respond to the return to the charity. To interpret the treatments, consider a simple social preference model building on DellaVigna *et al*. (2015) which embeds pure altruism and a version of "warm glow". The optimal effort is

$$e^*_{CH} = c'^{-1}(s + \alpha p_{CH} + a * .01). \tag{5}$$

In the simple, additive version of a pure altruism model à la Becker (1974), the worker cares about each dollar raised for the charity; as such, the altruism parameter α multiplies the return to the charity p_{CH} (equal to 0.01 or 0.10). In an alternative model, which we label "warm glow" (Andreoni, 1989), the worker still feels good for helping the charity, but she does not pay attention to the actual return to the charity; she just receives utility a for each button press to capture a warm glow or social norm of generosity.[10]

10. We use 'warm glow' to indicate the fact that workers feel good about the contribution to charity, but irrespective of the actual return to the charity. This warm glow specification, which is parallel to DellaVigna *et al*. (2015), is not part of the pre-registration. Noice that we multiply the warm glow parameter a by .01 (the return in the 1-cent treatment), without loss of generality, to facilitate the comparison between the two social preference parameters. Without rescaling, the estimates for a would be rescaled by 1/100.

The final social preference treatment is a gift exchange treatment modelled upon Gneezy and List (2006): "*In appreciation to you for performing this task, you will be paid a bonus of 40 cents. Your score will not affect your payment in any way.*" In this treatment there is no piece rate, but the "gift" may increase the motivation s by a factor Δs_{GE} reflecting reciprocity towards the employer.[11] Thus, the gift exchange effort equals

$$e^*_{GE} = c'^{-1}(s + \Delta s_{GE}). \tag{6}$$

2.4. *Time preferences*

Next, we have two discounting treatments: "*As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account two weeks from today*" and "*As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account four weeks from today.*" The piece rate is 1 cent as in a benchmark treatment, but the payment is delayed from nearly immediate ("*within 24 hours*") in the benchmark treatments, to 2 or 4 weeks later. This corresponds to the commonly-used experimental questions to capture present bias (Laibson, 1997; O'Donoghue and Rabin, 1999; Frederick *et al*., 2002).

We model the treatments with delayed payment with a present bias model:

$$e^*_t = c'^{-1}\left(s + \beta\delta^t p\right), \tag{7}$$

where β is the short-run impatience factor and δ is the long-run discounting factor. By comparing e^*_t in the discounting treatments to e^* in the piece rate treatments it is possible to back out the present bias parameter β and the (weekly) discounting factor δ.

An important caveat is that present bias should apply to the utility of consumption and real effort, not to the monetary payments per se, since such payments can be consumed in different periods (Augenblick *et al*., 2015). Having said this, the elicitation of present bias using monetary payments is very common.

2.5. *Reference dependence*

Next, we introduce treatments motivated by prospect theory (Kahneman and Tversky, 1979). A cornerstone of prospect theory is loss aversion: losses loom larger than gains. To measure loss aversion, we use a framing manipulation, as in Hossain and List (2012) and Fryer *et al*. (2012). The first treatment promises a 40-cent bonus for achieving a threshold performance: "*As a bonus, you will be paid an extra 40 cents if you score at least 2,000 points. This bonus will be paid to your account within 24 hours.* " The second treatment promises a 40 cent bonus, but then stresses that this payment will be lost if the person does not attain a threshold score: "*As a bonus, you will be paid an extra 40 cents. This bonus will be paid to your account within 24 hours. However, you will lose this bonus (it will not be placed in your account) unless you score at least 2,000 points.*" The payoffs are equivalent in the two cases, but the framing of the bonus differs. A third treatment is also on the gain side, for a larger 80-cent payment: "*As a bonus, you will be paid an extra 80 cents if you score at least 2,000 points. This bonus will be paid to your account within 24 hours.*"

11. The experiments on gift exchange in the field are motivated by laboratory experiments on gift exchange and reciprocity (Fehr *et al*., 1993; Fehr and Gächter, 2000).

For the gain treatments, subjects can earn payment G ($0.40 or $0.80) if they exceed a target performance T. Following the Koszegi-Rabin (2006) gain-loss notation (but with a reference point given by the status quo), the decision-maker maximizes

$$\max_{e\geq 0} se+\mathbf{1}_{\{e\geq T\}}G+\eta\left(\mathbf{1}_{\{e\geq T\}}G-0\right)-c(e). \tag{8}$$

The first term, $se+\mathbf{1}_{\{e\geq T\}}G$, captures the "consumption" utility, while the second term, $\eta(\mathbf{1}_{\{e\geq T\}}G-0)$, captures the gain utility relative to the reference point of no bonus. In the loss treatment, the decision-maker takes bonus G as reference point and thus maximizes

$$\max_{e\geq 0} se+\mathbf{1}_{\{e\geq T\}}G+\eta\lambda\left(0-\mathbf{1}_{\{e<T\}}G\right)-c(e). \tag{9}$$

The incentive to reach the threshold T is $(1+\eta)G$ in the gain condition versus $(1+\lambda\eta)G$ in the loss condition. Thus, with $\lambda>1$ (loss aversion) effort is higher in the loss treatment. The gain condition for $G=\$0.80$ has the purpose of benchmarking loss aversion: as we show in Section 7, observing effort in the three treatments allows us to identify the implied loss aversion λ (under the standard assumption $\eta=1$).[12]

A second key component of prospect theory is probability weighting: probabilities are transformed with a probability weighting function $\pi(P)$ which overweights small probabilities and underweights large probabilities (*e.g.* Prelec, 1998; Wu and Gonzalez, 1996). This motivates two treatments with stochastic piece rates, with expected incentives equal to the 1-cent benchmark treatment: "*As a bonus, you will have a 1% chance of being paid an extra $1 for every 100 points that you score. One out of every 100 participants who perform this task will be randomly chosen to be paid this reward.*" and "*As a bonus, you will have a 50% chance of being paid an extra 2 cents for every 100 points that you score. One out of two participants who perform this task will be randomly chosen to be paid this reward."*

In these treatments, the subjects earn piece rate p with probability P, and no piece rate otherwise, with $p*P=0.01$. The utility maximization is $\max_{e\geq 0} se+\pi(P)u(p)e-c(e)$, where $u(p)$ is the (possibly concave) utility of payment with $u(0)=0$. The effort e^* is

$$e^*_{PW,P}=c'^{-1}(s+\pi(P)u(p)). \tag{10}$$

A probability weighting function with prospect theory features implies $\pi(0.01)\gg 0.01$ and $\pi(0.5)<0.5$.[13] Thus, for $u(p)$ approximately linear, effort will be highest in the condition with 0.01 probability of a $1 piece rate: $e^*_{PW,P=.01}\gg e^*_{.01}>e^*_{PW,P=.5}$. Conversely, with no probability weighting and concave utility, the order is partially reversed: $e^*_{PW,P=0.01}<e^*_{PW,P=0.5}<e^*_{0.01}$.

2.6. *Psychology-based treatments*

A classical literature in psychology recognizes that human motivation is based to some degree on social comparisons (*e.g.* Maslow, 1943). Robert Cialdini has used comparisons to the achievements of others to induce motivation (*e.g.* Schultz *et al.*, 2007). In the ideal implementation, we would have informed the workers that a large majority of participants attain

12. To our knowledge, this is the first paper to propose this third condition, which allows for a simple measure of the loss aversion parameter λ.

13. In Section 6 we document that a meta-analysis of estimates of probability weighting implies $\pi(0.01)=0.06$ and $\pi(0.5)=0.45$.

a high threshold (such as 2,000 points). Given that we only report truthful messages, we opted for: "*Your score will not affect your payment in any way. Previously, many participants were able to score more than 2,000 points.*"[14]

A second social-comparison treatment levers the competitiveness of humans (*e.g.* Frank, 1985 within economics): "*Your score will not affect your payment in any way. After you play, we will show you how well you did relative to other participants.*"

The final manipulation is based on the influential literature in psychology on task significance (Grant, 2008): workers work harder when they are informed about the significance of their job. Within our setting, we inform people that "*Your score will not affect your payment in any way. We are interested in how fast people choose to press digits and we would like you to do your very best. So please try as hard as you can.*"

We model these psychological treatments as in (6) with a shift Δs in the motivation.

3. EXPERIMENT AND SURVEY DESIGN

Design logic. We designed the experiment with a dual purpose. First, we wanted to obtain evidence on behavioural motivators, covering present-biased preferences, reference dependence, and social preferences, three cornerstones of behavioural economics (Rabin, 1998; DellaVigna, 2009; Koszegi, 2014), as well as motivators borrowed more directly from psychology.

Second, we wanted to examine how experts forecast the impact of the various motivators. From this stand-point, we had five desiderata: (1) the experiment should have multiple treatments, to make the forecasting more informative; (2) the sample size for each treatment had to be large enough to limit the role for sampling variation, since we did not want the experts to worry about the precision of the estimates; (3) the differences in treatments had to be explained concisely and effectively, to give experts the best chance to grasp the design; (4) the results should be available soon enough, so that the experts could receive timely feedback; and (5) the treatments and forecasting procedure should be disclosed to avoid the perception that the experiments were selected on some criterion, that is, ones with counterintuitive results.

In light of this, we settled on a between-subject real-effort experiment run on Amazon Mechanical Turk (MTurk). MTurk is an online platform that allows researchers and businesses to post small tasks (referred to as HITs) that require a human to perform. Potential workers can browse the set of postings and choose to complete any task for the amount of money offered. MTurk has become very popular for experimental research in marketing and psychology (Paolacci and Chandler, 2014) and is also used increasingly in economics, for example for the study of preferences about redistribution (Kuziemko *et al*., 2015).

The limited cost per subject and large available population on MTurk allow us to run several treatments, each with a large sample size, achieving goals (1) and (2). Furthermore, the MTurk setting allows for a simple and transparent design (goal (3)): the experts can sample the task and can easily compare the different treatments, since the instructions for the various treatments differ essentially in only one paragraph. The MTurk platform also ensures a speedy data collection effort (goal (4)). Finally, we pre-registered both the experimental design and the survey, including a pre-analysis plan, to achieve goal (5).

14. We acknowledge that a number other than 2,000 could have been used as the social norm and a different norm may lead to more or less effort. This should be taken into consideration when thinking about the effectiveness of this treatment relative to the other treatments.

3.1. *Real-effort experiment*

With this framework in mind, we designed a simple real-effort task on MTurk. The task involved alternating presses of "a" and "b" for 10 minutes, achieving a point for each a-b alternation, a task similar to those used in the literature (Amir and Ariely, 2008; Berger and Pope, 2011). While the task is not meaningful per se, it does have features that parallel clerical jobs: it involves repetition and it gets tiring, thus testing the motivation of the workers. It is also simple to explain to both subjects and experts.

To enroll, the subjects go through three screens: (1) a recruiting screen, specifying a $1 pay for participating in an "*academic study regarding performance in a simple task*"[15], (2) a consent form, and (3) a page where they enter their MTurk ID and answer three demographic questions. The fourth screen provides instructions: "*On the next page you will play a simple button-pressing task. The object of this task is to alternately press the "a" and "b" buttons on your keyboard as quickly as possible for 10 minutes. Every time you successfully press the "a" and then the "b" button, you will receive a point. Note that points will only be rewarded when you alternate button pushes: just pressing the "a" or "b" button without alternating between the two will not result in points. Buttons must be pressed by hand only (key-bindings or automated button-pushing programs/scripts cannot be used) or the task will not be approved. Feel free to score as many points as you can.*" Then, the participant sees a different final paragraph (bold and underlined) depending on the condition to which they were randomly assigned. For example, in the 10-cent treatment, the sentence reads "*As a bonus, you will be paid an extra 10 cents for every 100 points that you score. This bonus will be paid to your account within 24 hours.*" Table 1 reports the key content of this paragraph for all eighteen treatments.[16] At the bottom of the page, subjects can try the task before proceeding.

On the fifth screen, subjects do the task. As they press digits, the page shows a clock with a 10-minute countdown, the current points, and any earnings accumulated (depending on the condition) (Online Appendix Figures 1 a–d). A sentence summarizes the condition for earning a bonus (if any) in that particular treatment. Thus, the eighteen treatments differ in only three ways: the main paragraph on the fourth screen explaining the condition, the one-line reminder in the task screen, and the rate at which earnings (if any) accumulate on the task screen.

After the 10 minutes are over, the subjects are presented with the total points, the bonus payout (if any) and the total payout, and can leave a comment if they wish. The subjects are then thanked for their participation and given a validation code to redeem their earnings.

3.1.1. Pre-registration. We pre-registered the design of the experiment on the AEA RCT Registry as AEARCTR-0000714 ("*Response of Output to Varying Incentive Structures on Amazon Turk*"). We pre-registered the rule for the sample size: we aimed to recruit 10,000 participants, and at least 5,000 participants based on a power study.[17] We ran the experiment for 3 weeks, at which point we had reached approximately 10,000 subjects.[18]

15. We require that workers have an 80% approval rate and at least fifty approved previous tasks.

16. For space reasons, in Table 1 we omit the sentence "*The bonus will be paid to your account within 24 hours.*" The sentence does not appear in the time discounting treatments.

17. Quoting from the registration, "*based on 393 pilot participants, the standard deviation of points scored was around 740 [...]. Assuming that this is approximately the standard deviation of each treatment in the experiment and [...] assuming [...] a sample size of 10,000 (555 per treatment), there is then an 80% power to reject the null hypothesis of zero difference when the actual difference is 124.6 points. Based on our pilot, different treatments can create differences in average points scored by as much as 400-500 points.*"

18. The registration documents states "*The task will be kept open on Amazon Mechanical Turk until either (i) two weeks have passed or (ii) 10,000 subjects have completed the study, whichever comes first. If two weeks pass without*

We also pre-specified the roles for sample inclusion: "*the final sample will exclude subjects that (1) do not complete the MTurk task within 30 minutes of starting or (2) exit then re-enter the task as a new subject (as these individuals might see multiple treatments) or (3) score 4000 or more points (as we have learned from a pilot study of* ∼300 *participants that it is physically impossible to score more than 3500 points, so it is likely that these individuals are using bots).*"

We ran the experiment before we collected forecasts so as to provide the experts with the results of three benchmark incentive treatments, thus conveying the curvature of the cost of effort function. At the same time, we wanted to ensure that there would be no leak of any results. As such, as authors we did not have access to experimental results until the end of the collection of the expert forecasts, in September 2015. During the MTurk experiment, a research assistant ran a script to monitor the sample size and the results in the three benchmark treatments, and sent us daily updates which we monitored for potential data issues.

3.1.2. Data collection. The experiment ran for 3 weeks in May 2015. The initial sample consists of 12,838 MTurk workers who started our experimental task. Of these, 721 were dropped because of a technical problem with the survey over a several-hour period when the software program Qualtrics moved to a new server. Individuals during this time period experienced a malfunctioning of the counter that kept track of their scores. This sample exclusion, which we could not have anticipated, does not appear in the registration.

We then applied the three specified sample restrictions. We dropped (1) 48 workers for scoring above 4,000 points, (2) 1,543 workers for failing to complete the experiment (*e.g.* many participants only filled out the demographics portion of the experiment and were never assigned a treatment), and (3) 364 workers for stopping the task and logging in again. (We stated in the instructions to the workers that they could not stop the task and log in again.) Two additional restrictions were added: we dropped 187 workers because their HIT was not approved for some reason (*e.g.* they did not have a valid MTurk ID) as well as 114 workers who never did a single button press. These participants may have experienced a technical malfunction or it may be that their results were not recorded for some reason.[19]

Many of the participants that dropped out of our study did so after seeing their treatment assignment. Thus, one may worry about selective attrition. A Pearson chi-squared test provides some evidence that the drop-out frequencies are not equal across treatments ($p = 0.034$). Still, the actual attrition is quite small and a simple calibration suggests that it cannot lead to a large change in effort levels across conditions. In addition, when it comes to the expert forecasts, any selective attrition should already be considered, given that we provide experts with the effort in three benchmark conditions (no pay, 1-cent, and 10-cent) for the non-attrited sample. Thus, the experts are calibrated with results that contain the selective attrition.

3.1.3. Summary statistics. The final sample includes 9,861 subjects, about 550 per treatment. As Online Appendix Table 1 shows, the recruited MTurk sample matches the U.S. population for gender, and somewhat over-represents high-education groups and younger

5500 subjects completing the task, then the task will be kept open (up to six weeks) until 5500 subjects are obtained." We deviated slightly from this rule by running the experiment for 3 weeks because we incorrectly thought that we registered a 3-week duration. The deviation has minor impact as (1) 80% of subjects had been recruited by the end of week 2, and (2) the authors did not monitor the experimental results during the three weeks (other than for the three benchmark conditions), thus removing the potential for selective stopping.

19. The two additional restrictions, which are immaterial for the results, were added before we analysed the full data and were included in the pre-registration for the survey protocol AEARCTR-0000731 (see below).

individuals. This is consistent with previous literature documenting that MTurkers are actually quite representative of the population of U.S. internet users (Ipeirotis, 2010; Ross *et al*., 2010; Paolacci, 2010) on characteristics such as age, socioeconomic status, and education levels.

3.2. *Expert Survey*

3.2.1. Survey. The survey of experts, registered as AEARCTR-0000731, is formatted with the platform Qualtrics and consists of two pages.[20] In the main page, the experts read a description of the task, including the exact wording seen by the MTurkers. The experts can experience the task by clicking on a link and see the screenshots viewed by the MTurk workers with another click. The experts are then informed of a prize that depends on the accuracy of their forecasts. "*Five people who complete this survey will be chosen at random to be paid [...] These five individuals will each receive $1,000 - (Mean Squared Error/200), where the mean squared error is the average of the squared differences between his/her answers and the actual scores.*" This structure is incentive compatible under risk neutrality: participants who minimize the sum of squared errors should indicate as their forecast the mean expected effort by treatment.[21]

The survey then displays the mean effort in the three benchmark treatments: no-piece rate, 1-cent, and 10-cent piece rate. The experts then see a list of the remaining fifteen treatments and create a forecast by moving the slider, or typing the forecast in a text box (though the latter method was not emphasized) (Online Appendix Figure 2). The experts can scroll back up on the page to review the instructions or the results of the benchmark treatments.[22]

We decided ex ante the rule for the slider scale. We wanted the slider to include the values for all eighteen treatments while at the same time minimizing the scope for confusion. Thus, we chose the minimum and maximum unit to be the closest multiple of 500 that is at least 200 units away from all treatment scores. A research assistant checked this rule against the results, leading to a slider scale between 1,000 and 2,500.

3.2.2. Experts. To form the group of behavioural experts, we form an initial list including: (1) authors of papers presented at the Stanford Institute of Theoretical Economics (SITE) in Psychology and Economics or in Experimental Economics from its inception until 2014 (for all years in which the program is online); (2) participants of the Behavioural Economics Annual Meeting (BEAM) conferences from 2009 to 2014; (3) individuals in the program committee and keynote speakers for the Behavioural Decision Research in Management Conference (BDRM) in 2010, 2012, and 2014; (4) invitees to the Russell Sage Foundation 2014 Workshop on "Behavioural Labor Economics", and (5) a list of behavioural economists compiled by ideas42. We also add by hand a small number of additional experts. We then pare down this list of over 600 people to 312 researchers to whom at least one of the two authors had some connection.

On July 10 and 11, 2015 one of the us sent a personalized email to each expert. The email provided a brief introduction and notified about an upcoming email from Qualtrics with a unique

20. We provide further details on the survey in DellaVigna and Pope (forthcoming).

21. We avoided a tournament payout structure (paying the top five performers) which could have introduced risk-taking incentives; we pay instead five randomly drawn participants.

22. To test for fatigue, we randomize across experts the order of the treatments (the only randomization in the survey). Namely, we designate six possible orders, always keeping related interventions together, to minimize the burden on the experts. There is no evidence of fatigue effects.

TABLE 2
Summary statistics, experts

	All experts contacted	Experts completed survey	Experts completed all fifteen treatments
	(1)	(2)	(3)
Primary field			
Behavioural Econ.	0.24	0.31	0.32
Behavioural Finance	0.06	0.05	0.04
Applied Micro	0.17	0.19	0.19
Economic Theory	0.09	0.07	0.07
Econ. Lab Exper.	0.17	0.15	0.16
Decision Making	0.17	0.12	0.12
Social Psychology	0.08	0.10	0.10
Academic rank			
Assistant Professor	0.26	0.36	0.36
Associate Professor	0.15	0.15	0.15
Professor	0.55	0.45	0.45
Other	0.04	0.04	0.04
Minutes spent (med.)			17
Clicked practice task			0.44
Clicked instructions			0.22
Heard of Mturk			0.98
Used Mturk			0.51
Observations	312	213	208

Notes: The Table presents summary information on the experts participating in the survey. Column (1) presents information on the experts contacted and Column (2) on the experts that completed the survey. Column (3) restricts the sample further to subjects who made a forecast for all fifteen treatments.

link to the survey. We followed up with an automated reminder email about 2 weeks later to experts who had not yet completed the survey (and had not expressed a desire to opt out from communication), and with a final personal email afterwards to the non-completers.[23]

Out of the 312 experts sent the survey, 213 completed it, for a participation rate of 68%. The main sample of 208 experts does not include five responses with missing forecasts for at least one of the fifteen treatments. Table 2 shows the selection into response. Notice that the identity of the respondents is kept anonymous. On 30, November 2015, each expert received a personalized email with a link to a figure analogous to Figure 5 that also included their own forecasts. We also drew winners and distributed the prizes as promised.

4. EFFORT BY TREATMENT

4.1. *Average effort*

4.1.1. Piece rate treatments. We start the analysis from the benchmark treatments which the experts had access to. Incentives have a powerful effect on effort, raising performance from an average of 1,521 points (no piece rate) to 2,029 (1-cent piece rate) and 2,175 (10-cent piece rate). The standard error for the mean effort per treatment is around 30 points or less (Table 3), implying that differences across treatments larger than 85 points are statistically significant.

23. We also collected forecasts from PhD students in economics, undergraduate students, MBA students, and a group of MTurk subjects. We analyse these results in DellaVigna and Pope (forthcoming).

TABLE 3
Findings by treatment: effort in experiment and expert forecasts

Category	Treatment wording	*N*	Mean effort (s.e.)	Mean forecast	Std. dev. forecast	Actual - forecast (s.e.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Piece rate	"Your score will not affect your payment in any way."	540	1521 (31.23)		Benchmark	
	As a bonus, you will be paid an extra **1 cent** for every **100 points** that you score."	558	2029 (27.47)		Benchmark	
	"As a bonus, you will be paid an extra **10 cents** for every **100 points** that you score."	566	2175 (24.28)		Benchmark	
	"As a bonus, you will be paid an extra **4 cents** for every **100 points** that you score."	562	2132 (26.42)	2057	120.87	75 (27.72)
Pay enough or don't pay	"As a bonus, you will be paid an extra **1 cent** for every **1,000 points** that you score."	538	1883 (28.61)	1657	262.00	226 (33.89)
Social preferences: charity	"As a bonus, the Red Cross charitable fund will be given **1 cent** for every 100 points that you score."	554	1907 (26.85)	1894	202.20	13 (30.29)
	"As a bonus, the Red Cross charitable fund will be given **10 cents** for every 100 points that you score."	549	1918 (25.93)	1997	196.75	−79 (29.30)
Social preferences: gift exchange	"In appreciation to you for performing this task, you will be paid a **bonus** of **40 cents**. Your score will not affect your payment in any way."	545	1602 (29.77)	1709	207.12	−107 (33.05)
Discounting	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **two weeks** from today."	544	2004 (27.38)	1933	142.03	71 (29.10)
	"As a bonus, you will be paid an extra 1 cent for every 100 points that you score. This bonus will be paid to your account **four weeks** from today."	550	1970 (28.68)	1895	162.54	75 (30.81)
Gains versus losses	"As a bonus, you will be paid an **extra 40 cents** if you score at least 2,000 points."	545	2136 (24.66)	1955	149.90	181 (26.76)
	"As a bonus, you will be paid an **extra 40 cents**. However, you will **lose this bonus** (it will not be placed in your account) unless you score at least 2,000 points. "	532	2155 (23.09)	2002	143.57	153 (25.14)
	"As a bonus, you will be paid an **extra 80 cents** if you score at least 2,000 points."	532	2188 (22.99)	2007	131.93	181 (24.74)
Risk aversion and probability weighting	"As a bonus, you will have a **1% chance** of being paid an **extra $1** for every 100 points that you score. One out of every 100 participants who perform this task will be randomly chosen to be paid this reward."	555	1896 (28.44)	1967	253.43	−71 (33.43)

(*continued*)

TABLE 3
Continued

Category	Treatment wording	*N*	Mean effort (s.e.)	Mean forecast	Std. dev. forecast	Actual - forecast (s.e.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
	"As a bonus, you will have a **50% chance** of being paid an **extra 2 cents** for every 100 points that you score. One out of two participants who perform this task will be randomly chosen to be paid this reward."	568	1977 (24.73)	1941	179.27	36 (27.68)
Social comparisons	"Your score will not affect your payment in any way. In a previous version of this task, **many participants** were able to score more than **2,000 points.**"	526	1848 (32.14)	1877	209.48	−29 (35.27)
Ranking	"Your score will not affect your payment in any way. After you play, we will show you **how well you did relative** to other participants who have previously done this task."	543	1761 (30.63)	1850	234.28	−89 (34.67)
Task significance	"Your score will not affect your payment in any way. We are interested in how fast people choose to press digits and we would like you to do your very best. So **please try as hard as you can**."	554	1740 (28.76)	1757	230.15	−17 (32.89)

Notes: The Table lists the eighteen treatments in the Mturk experiment. The treatments differ just in one paragraph explaining the task and in the vizualization of the points earned. Column (2) reports the key part of the wording of the paragraph. For brevity, we omit from the description the sentence "This bonus will be paid to your account within 24 hours" which applies to all treatments with incentives other than in the Time Preference ones where the payment is delayed. Notice that the bolding is for the benefit of the reader of the Table. In the actual description to the MTurk workers, the whole paragraph was bolded and underlined. Column (1) reports the conceptual grouping of the treaments, Columns (3) and (4) report the number of MTurk subjects in that treatment and the mean number of points, with the standard errors. Column (5) reports the mean forecast among the 208 experts of the points in that treatment. Column (6) reports the standard deviation among the expert forecasts for that treatment. Column (7) reports the difference between the average forecast and the actual average effort, with its standard errror.

Using as moments the average effort in these benchmark treatments, we estimate the cost function using a minimum distance estimator. The model which we pre-registered assumes a power cost function, leading to expression (2) for effort e^*. We estimate the three parameters: the motivation s, the cost curvature (and inverse of the elasticity) γ, and the scaling parameter k. Hence, we are exactly identified with three moments and three parameters.

As Column 1 of Table 5 shows,[24] the cost of effort has a high estimated curvature ($\hat{\gamma} = 33$) and thus a low elasticity of 0.03. This is not surprising given that an order-of-magnitude increase in the piece rate (from 1 to 10 cents) increases effort by less than 10%. The estimated motivation $\hat{s}$ is very small: given the high curvature of the cost of effort function, even a small degree of motivation can reproduce the observed effort of 1,522 for zero piece rate.

24. The standard errors for the parameters are derived via a bootstrap with 1,000 draws.

How does this estimated model fit in sample (the benchmark treatments) and out of sample (the 4-cent piece rate)? Figure 2a displays the estimated marginal cost curve $c'(e)=\hat{k}e^{\hat{\gamma}}$ and the marginal benefit curves $\hat{s}+p$ for the different piece rates. By design, the model perfectly fits in sample the 0-cent, 1-cent, and 10-cent cases. The model then predicts a productivity for the 4-cent case of 2,116, very close to the actual effort of 2,132.

As an alternative cost of effort function, as discussed in Section 2, we consider an exponential function, with declining elasticity: $c(e)=k\exp(\gamma e)/\gamma$. Column 3 of Table 5 shows that, as with the power function, the motivation s is estimated to be very small. The exponential function also perfectly fits the benchmark moments, and makes a similar prediction for the 4-cent treatment (Online Appendix Figure 3 a). Further, allowing for heterogeneity and discrete incentives also leads to a very similar prediction of effort (Section 7).

4.1.2. Pay enough or don't pay at all. In the first behavioural treatment we pay a very low piece rate: 1 cent for every 1,000 points. For comparison, the 1-cent benchmark treatment pays 1 cent per 100 points, and thus has *ten* times higher incentives. We examine whether this very low piece rate crowds out motivation as in Gneezy and Rustichini (2000).

To estimate the extent of crowd-out, we predict the counterfactual effort given the incentive, assuming no crowd-out (*i.e.* zero Δs_{CO} in expression (4)): $\hat{e}_{CO}=\left((\hat{s}+0.001)/\hat{k}\right)^{1/\hat{\gamma}}$.[25] Figure 2b displays the predicted effort, 1,893, at the intersection of the marginal cost curve with the marginal benefit set at $\hat{s}+0.001$. The model with exponential cost of effort makes a very similar prediction (Online Appendix Figure 3 b), as do models allowing for heterogeneity and discrete incentives (see Section 7 and Online Appendix A). Remarkably, the observed effort, 1,883, equals almost exactly the predicted effort due to incentives. The very low piece rate did not crowd out motivation in our setting.

4.1.3. Social preferences. Next, we consider the two charitable giving treatments, in which the Red Cross receives 1 cent (or 10 cents) per 100 points. Figure 3 shows the average effort for all eighteen treatments, ranked by average effort. The 1-cent charity treatment induces effort of 1,907, well above the no-piece rate benchmark, but below the treatment with a private 1-cent piece rate. This indicates social preferences with a smaller weight on a charity than on oneself. Interestingly, the 10-cent charity treatment induces almost identical effort, 1,918, suggesting that individuals are not responsive to the return to the charity.

The third social preference treatment involves gift exchange: subjects receive an unexpected bonus of 40 cents, *unconditional* on performance. As Figure 3 and Table 3 show, this treatment, while increasing output relative to the no-pay treatment, has the second smallest effect, 1,602, after the benchmark no-piece-rate treatment.

4.1.4. Time preferences. The two time preference treatments mirror the 1-cent benchmark treatment, except that the promised amount is paid in 2 (or 4) weeks. Figure 3 shows that the temporal delay in the payment lowers effort somewhat, but the effect is quantitatively quite small. More importantly, we do not appear to find evidence for a beta-delta pattern: if anything, the decline in output is larger going from the 2-week treatment to the 4-week treatment than from the immediate pay to the two-week payment.

25. As piece rate we use one tenth the piece rate for the benchmark one-cent treatment ($p=0.01$), ignoring the fact that the piece rate paid only every 1,000 points. We return to this later in Appendix A.

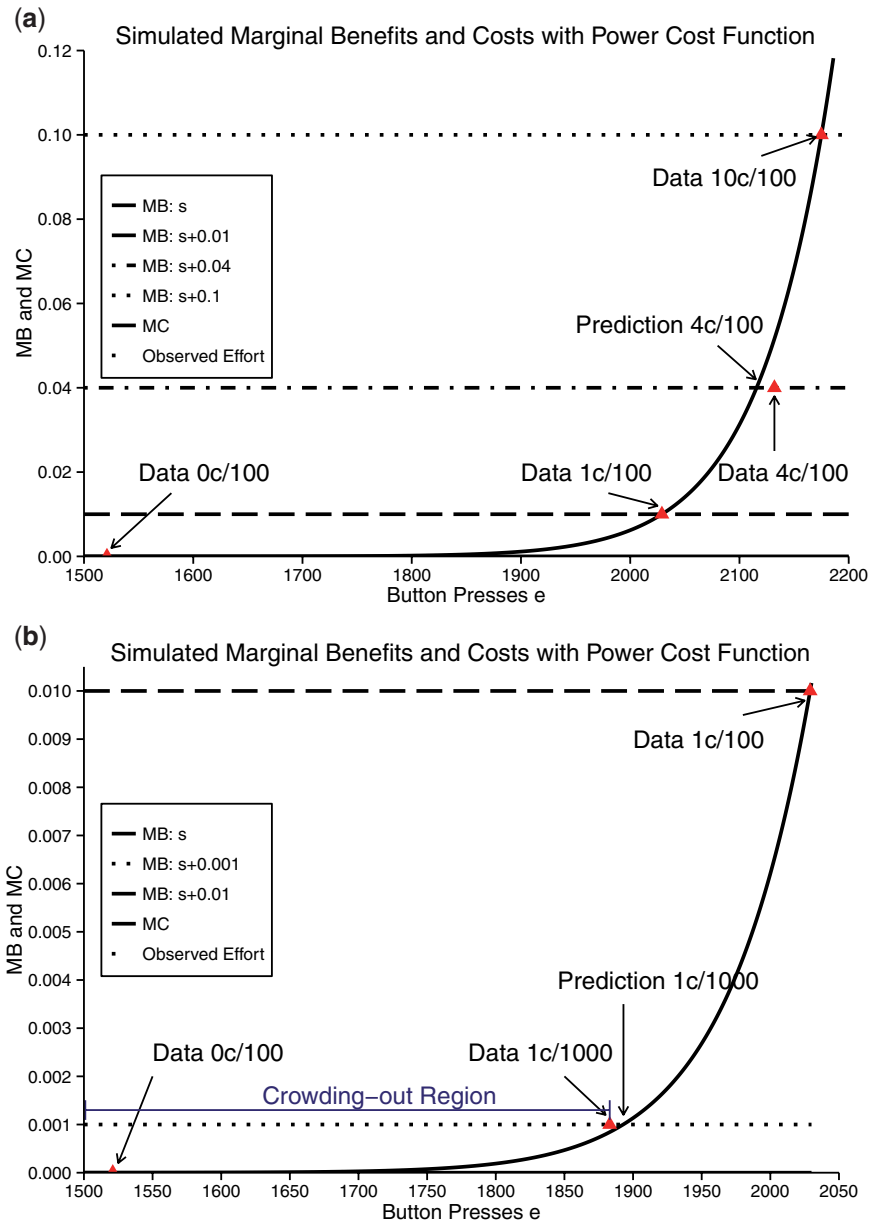

FIGURE 2

Estimate of model on three benchmark treatments. (a) Estimate with 0c, 1c, 10c piece rate and prediction for 4c piece rate. (b) Predicted effort for "Paying Too Little" treatment (1 cent for 1,000 presses)

Notes: Figure 2a plots the marginal cost curve and the marginal benefit curve for the three benchmark treatments for the power cost function estimates. The marginal benefit curve equals the estimated s (warm glow) plus the piece rate. The marginal cost curve equals ke^s at the estimated k and s. At the estimates, we fit the three benchmark levels of effort perfectly, given that the model is just identified. Figure 2a also plots the out of sample prediction for the 4 cent treatment (which is not used in the estimates), as well as the observed effort for that treatment. Figure 2b plots, for the same point estimates, the out of sample prediction for the treatment with 1-cent per 1,000 clicks.

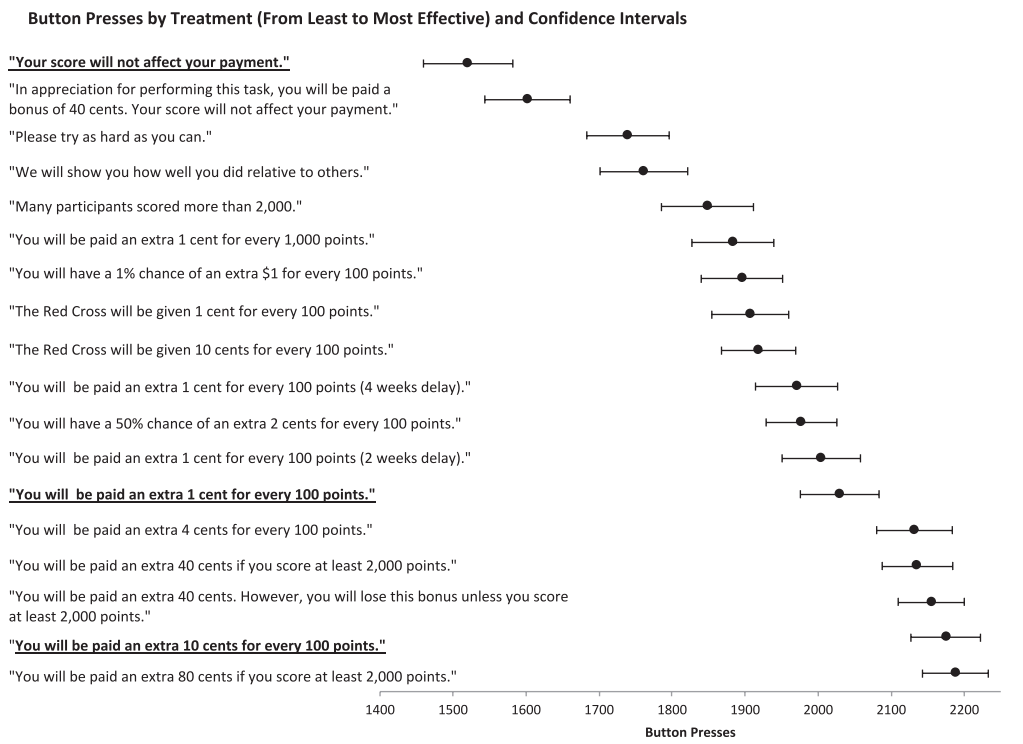

FIGURE 3

Average button presses by treatment in Amazon Turk task

Notes: Figure 3 presents the average score and confidence interval for each of eighteen treatments in a real-effort task on Amazon Turk. Participants in the task earn a point for each alternating a-b button press within a 10-minute period. The eighteen treatments differ only in one paragraph presenting the treatments, the key sentence of which is reproduced in the first row. Each treatment has about 550 participants.

4.1.5. Reference dependence. Next, we focus on loss aversion with treatments that vary the framing of a bonus at a 2,000 threshold as a gain or loss. As Figure 3 shows, the effort is higher for the 40-cent loss framing than for the 40-cent gain framing, though the difference is small and not statistically significant. In terms of induced output, the 40-cent loss treatment is about halfway between the 40-cent gain treatment and the 80-cent gain treatment. We return in Section 7 to the implied loss aversion coefficient.

Another key component of reference dependence is the probability weighting function which magnifies small probabilities. We designed two treatments with stochastic piece rates yielding (in expected value) the same incentive as the 1-cent benchmark: a treatment with 1% probability of a $1 piece rate (per 100 points) and another with 50% probability of a 2c piece rate (also per 100 points). Under probability weighting (and approximate risk neutrality), the 1-percent treatment should have the largest effect, even compared to the 1-cent benchmark. We find no support for overweighting of small probabilities: the treatment with 1% probability of $1 yields significantly lower effort (1,896) compared to the benchmark 1-cent treatment (2,029) or the 50-percent treatment (1,977).

4.1.6. Psychology-based treatments. Finally, we turn to the more psychology-motivated treatments, which offer purely non-monetary encouragements: social comparisons

(Schultz *et al*., 2007), ranking with other participants, and emphasis of task significance (Grant, 2008).

All three treatments outperform the benchmark no-piece-rate treatment by 200–300 points, with the most effective treatment being the Cialdini-base social comparison. The treatments also are more effective than the (equally unincentivized) gift-exchange treatment. At the same time, they are less effective than any of the treatments with incentives, including even the very-low-pay treatment. At least in this particular task with MTurk workers, purely psychological interventions have only a moderate effectiveness relative to the power of incentives. Still, they are cost-effective as they increase output for no additional cost.

4.2. *Heterogeneity and timing of effort*

4.2.1. Distribution of effort. Beyond the average effort, which is the variable that the experts forecast, we consider the distribution of effort (Online Appendix Figure 4). Across all eighteen treatments, relatively few workers do fewer than 500 presses, and even fewer score more than 3,000 points with almost no one above 3,500 points. There are spikes at each 100 and especially at each 1,000-point mark, in part because of discrete incentives at these round numbers.

Figure 4a presents the cumulative distribution function for the benchmark treatments and for the crowd-out treatment.[26] Incentives induce a clear rightward shift in effort relative to the no-pay benchmark, even with the very low 1-cent per 1,000-points piece rate. The piece rates are particularly effective at reducing the incidence of effort below 1,000 points, from 20% in the no-pay benchmark to less than 8% in any of the piece rate conditions.

Figure 4b shows that the treatments with no monetary incentives shift effort to the right, though not as much as the piece rate treatments do. Despite the absence of monetary incentives, there is some evidence of bunching at round numbers of points.

Regarding the gain-loss treatments (Figure 4c), we observe, as expected, bunching at 2,000 points, the threshold level for earning the bonus, and missing mass to the left of 2,000 points. Compared to the 40-cent gain treatment, both the 80-cent gain and the 40-cent loss treatments have 5% less mass to the left of 2,000 points, and more mass at 2,000 points (the predicted bunching) and points in the low 2,000s. The difference between the three treatments is smaller for low effort (below 1,500 points) or for high effort (above 2,500 points).[27] This conforms to the model predictions: individuals who are not going to come close to 2,000 points, or individuals who were planning to work hard nonetheless, are largely unaffected by the incentive change. These findings are in line with evidence on bunching and shifts due to discrete incentives and loss aversion (*e.g*. Rees-Jones, 2014; Allen *et al*., 2017).[28]

4.2.2. Effort over time. As final piece of evidence on the MTurker effort, Online Appendix Figure 5a and b display the evolution of effort over the 10 minutes of the task.

26. The c.d.f. of effort for the 4-cent treatment, which would be hard to see in the figure, lies between the 1-cent and the 10-cent benchmarks.

27. Formally, there should be no impact of the change in incentive on the distribution of points about 2,000. However, some small slippage from the threshold at 2,000 is natural.

28. A comparison with the no-piece rate benchmark also shows that the threshold incentive doubles the share of workers exerting effort above 2,500 points. This difference is not predicted by a simple reference-dependence model, given that there is no incentive to exert effort past the 2,000-point threshold. For the estimation of reference dependence, we compare the three threshold treatments to each other and thus do not take a stand on the level of effort induced by the threshold itself.

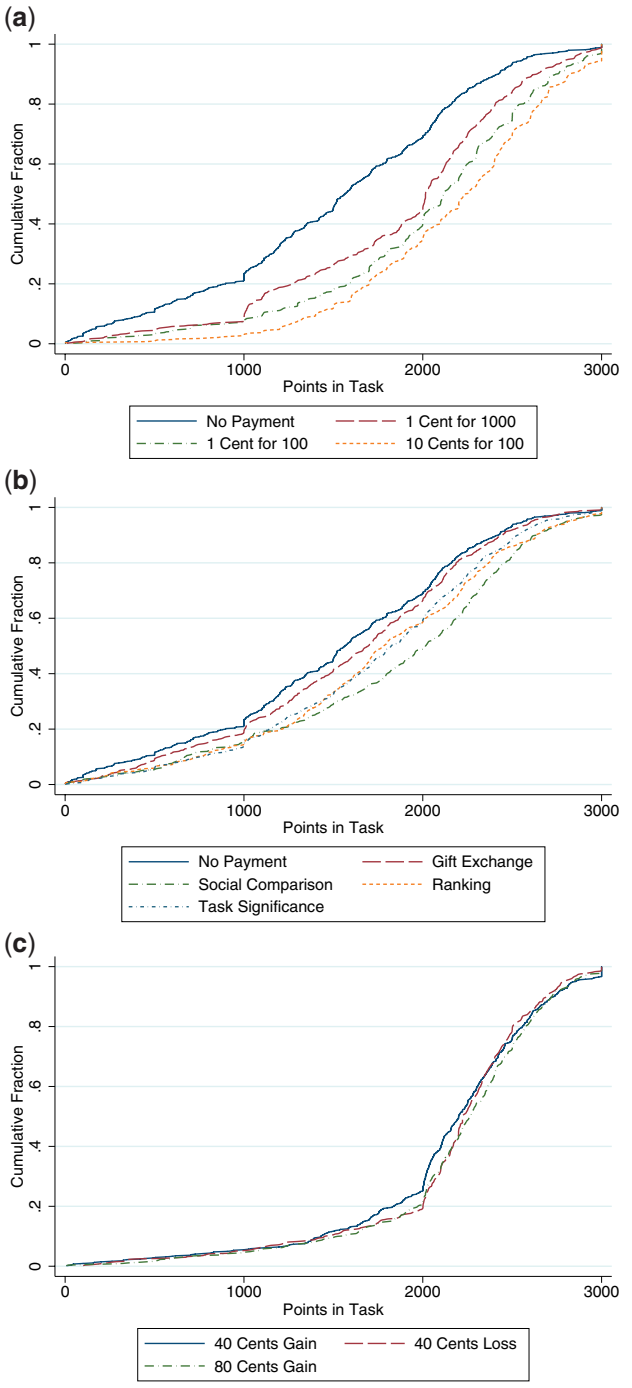

FIGURE 4

Distribution of effort, MTurk Workers, cumulative distribution function. (a) Piece-rate treatments. (b) Treatments with no monetary payoff. (c) Gain–loss treatments

Notes: Figures 4a–c present the cumulative distribution function of points for the MTurk workers in each of the treatments featured. The sample size in each treatment is approximately 550 subjects. Figure 4a features the three benchmark treatments (no piece rate, 1-cent per 100 points and 10 cents per 100 points), as well as the low-piece-rate treatment, 1 cent per 1,000 points. Figure 4b presents the results for the four treatments with no incentives (except for the charity treatments). Figure 4c presents the results for the gain–loss treatments.

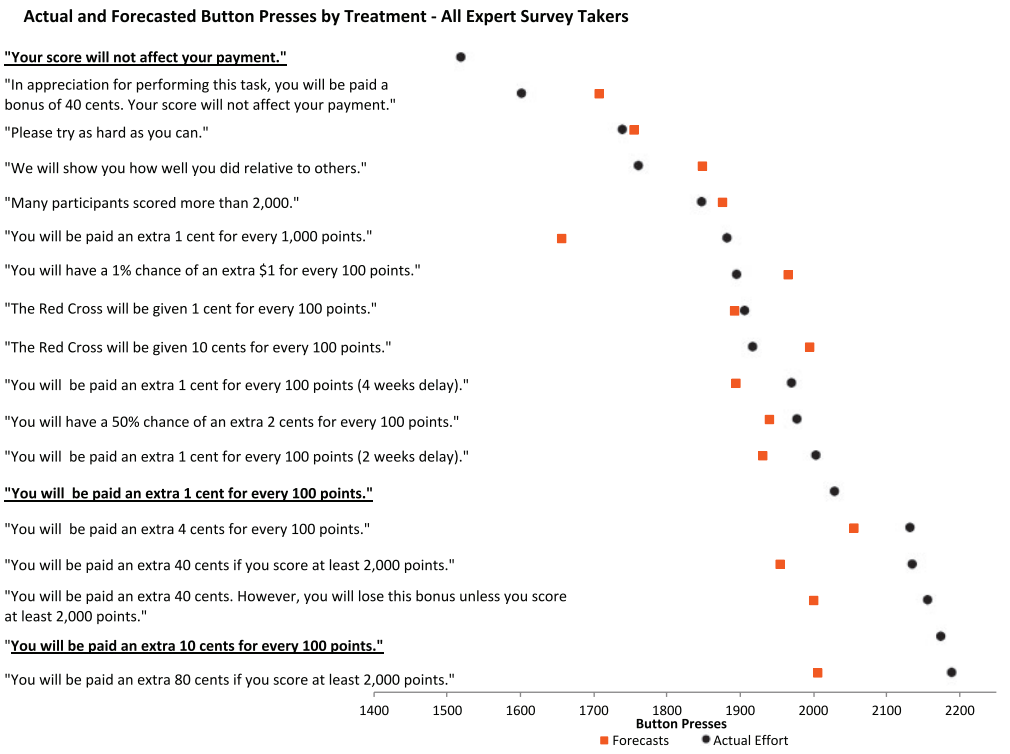

FIGURE 5

Average button presses by treatment and average expert forecasts

Notes: The black circles in Figure 5 present the average score for each of eighteen treatments in a real-effort task on Amazon Turk. Participants in the task earn a point for each alternating a-b button press within a 10-minute period. The eighteen treatments differ only in one paragraph presenting the treatments, the key sentence of which is reproduced in the first row. Each treatment has about 550 participants. The squares represent the average forecast from the sample of 208 experts who provided forecasts for the treatments. The three bolded treatments are benchmarks; the average score in the three benchmarks was revealed to the experts and thus there is no forecast.

Overall, the average effort remains relatively constant, potentially reflecting a combination of fatigue and learning by doing. The only treatments that, not surprisingly, experience a substantial decrease of effort in the last 3 minutes are the gain/loss treatments, since the workers are likely to have reached the 2,000 threshold by then. The plots also show a remarkable stability in the ranking of the treatments over the different minutes: for example, at any given minute, the piece rate treatments induce a higher effort than the treatments with non-monetary pay. The one exception is the crowd-out treatment which in the final minutes declines in effectiveness.

5. EXPERT FORECASTS

5.1. *Mean expert forecasts*

Which of these results did the experts anticipate? What are the biggest discrepancies? For each treatment, Figure 5 and Table 3 indicate the mean forecast across the 208 experts, along with the actual effort. Table 3 also indicates whether there is a statistically significant difference between the mean forecast and the effort.

The largest discrepancy (more than 200 points) between mean forecast and effort is for the low-pay treatment: on average, experts expect crowd out with a very low piece rate, at least with respect to the counterfactual computed above. Instead, we find no evidence of crowd out.

The next largest deviations occur for the gain–loss treatments: experts expect these treatments to induce an effort of around 2,000 points while the observed effort is around 2,150 points. Notice that this deviation reflects an incorrect expectation regarding the effect of the threshold, *not* a discrepancy about the gain–loss framing. Regarding the latter, the forecasters on average expect about the same effort from the 80-cent gain treatment (2,007) and from the 40-cent loss treatment (2,002). We return to this in Section 7.

Another sizeable deviation is for the gift exchange treatment which, as we noted, has a very limited effect on productivity. Forecasters on average expect an impact of gift exchange that is 107 points larger, 1,709 points versus 1,602 points.

Turning to the charitable giving treatments, the experts are spot on (on average) with their forecast for the 1-cent charitable giving treatment, 1,894 versus 1,907 points. They however predict that the 10-cent charitable giving treatment will yield output that is about 80 points higher, whereas the output is essentially the same under the two conditions. The forecasters expect pure altruism to play a role, while the evidence points almost exclusively to warm glow. We decompose formally the two components in Section 7.

It is interesting to consider together all the six treatments with no private monetary incentives: gift exchange, the psychology-based treatments, and the charitable-giving treatments. The experts are remarkably accurate: the average forecast ranks the six treatments in the exact correct order of effectiveness, from gift exchange (least effective) to 10-cent charitable giving (most effective). Furthermore, the deviation between average forecast and actual performance is at most 107 points, a deviation of less than 7% from the actual effort.

Considering then the time preference treatments, the experts expect a significant output decrease with a 2-week delay, compared to the 1-cent treatment with no delay, with only a small further decrease for a 4-week delay. The experts thus anticipate present bias, while the evidence is more consistent with delta discounting. We return to this in Section 7.

Finally in the treatments with probabilistic piece rate, the experts on average guess just right the output for the treatment with a 50% probability of a 2-cent piece rate (1,941 versus 1,977). However, they on average expect that the effort will be somewhat higher for the treatment with a 1% chance of a $1 piece rate, in the direction predicted by probability weighting (though with a modest magnitude). The evidence, instead, does not support the overweighting of small probabilities predicted by probability weighting.

5.2. *Heterogeneity of expert forecasts*

How much do experts disagree? We consider the dispersion of forecasts in Figures 6a-d, displaying also the observed average effort (circles) and the benchmarks (vertical lines).

Two piece rate treatments are polar opposites in terms of expert disagreement (Figure 6a). The 4-cent treatment has the least heterogeneity in forecasts, not surprisingly since one can form a forecast using a straightforward model. In contrast, the 1-cent per 1,000-point treatment has the most heterogeneity. About 35% of experts expect strong enough motivational crowd out to yield lower output relative to the no-pay treatment (the first vertical line), while other experts expect no crowd out.

The forecasts for the charity treatments (also in Figure 6a) also display a fair degree of disagreement on the expected effectiveness: 20% of experts expects the 1-cent charity treatment to outperform the 1-cent piece rate treatment. These experts expect that workers assign a higher

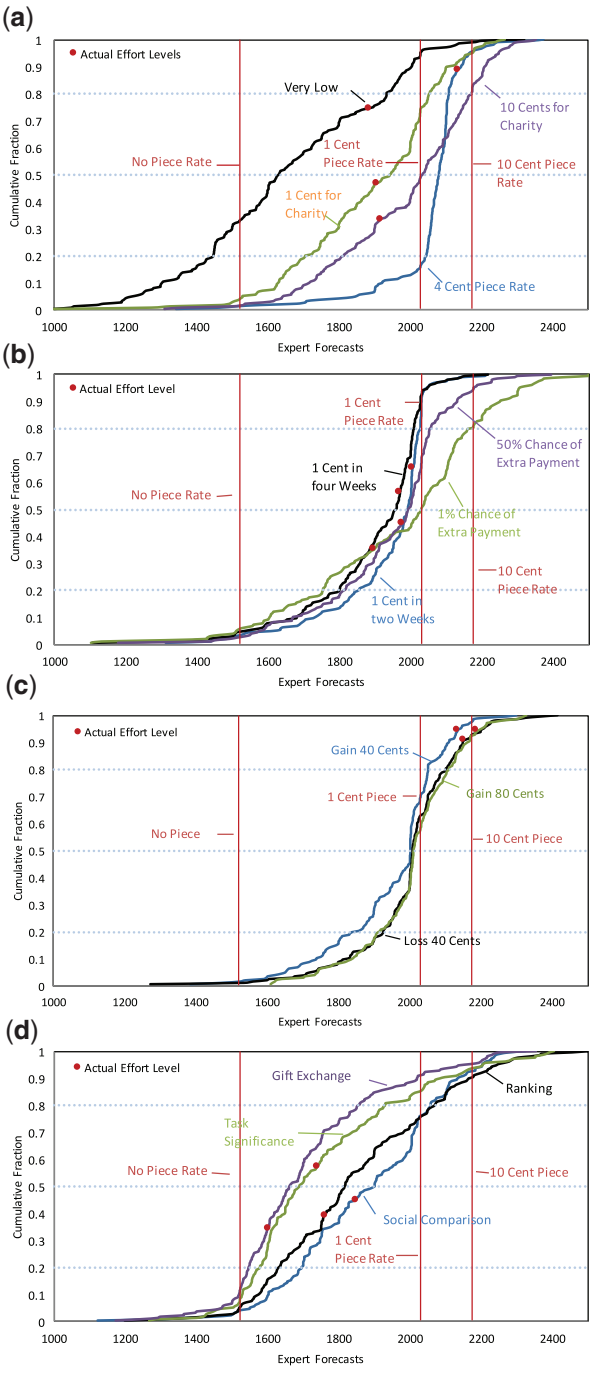

FIGURE 6

Heterogeneity of expert forecasts, cumulative distribution function. (a) Piece-rate and charity treatments. (b) Time preference and probability weighting treatments. (c) Gain and loss treatments. (d) Gift exchange and psychology treatments

Notes: Figures 6a–d present the cumulative distribution function of forecasts by the 208 experts (see Table 1 for the list of treatments). The circles present the actual average score for each treatment. The vertical lines present the score in the three benchmark treatments. Since the average score in the three benchmarks was revealed to the experts, there is no forecast for those.

weight on the return to a charity than on an equal-size private return. The disagreement is instead limited for the delayed-payment treatments (Figure 6b).

The probability weighting treatments (also in Figure 6b) reveal substantial heterogeneity. Fifty percent of experts expect higher effort in the 1 percent treatment than in the 1-cent benchmark; of these experts, almost half expects strong enough overweighting of small probabilities to lead to higher effort than in the 10-cent benchmark. The remaining 50% of experts instead expects risk aversion (over small stakes) to be a stronger force. There is much less variance among experts for the 50% treatment, as one would expect, since probability weighting, to a first approximation, should not play a role.

Figure 6c presents the evidence for the gain and loss treatments, showing that the c.d.f.s for the 80-cent gain and the 40-cent loss treatment are right on top of each other.

For the remaining treatments with no incentive pay—gift exchange and the psychology treatments—there is a fairly wide distribution of guesses mostly between the no-pay treatment and the 1-cent piece rate treatment (Figure 6d). For the two social comparison treatments, in fact, 25% of experts expect that these treatments would outperform the 1-cent piece rate treatment. In reality, the treatments, while effective, are not that powerful.

5.2.1. Field. Is the heterogeneity in forecasts explained in part by differences in the field of expertise? Figure 7 presents the average forecast by treatment separately for experts with primary field in behavioural economics, laboratory experiments, standard economics, and psychology and decision-making. Perhaps surprisingly, the differences are small. All groups of experts expect more crowd out than in the data, expect more gift exchange than in the data, and expect higher effort for the 10-cent charitable giving treatment compared to the 1-cent charitable giving treatment. There are some differences—psych experts expect less overweighting of small probabilities—but the differences are small and unsystematic. Field of expertise, thus, does not explain the heterogeneity in forecasts.[29]

6. INTERPRETATION AND META-ANALYSIS

How do we interpret the differences between the experimental results and the expert forecasts? We consider three classes of explanations: *biased literature*, *biased context*, and *biased experts*. In the first explanation, *biased literature*, the published literature upon which the experts rely is biased, perhaps due to its sparsity or some form of publication bias. In the second explanation, *biased context*, the literature itself is not biased, but our experimental results are unusual and differ from the literature due to our particular task or the subject pool in our study.[30] In this explanation, experts may be unable to fully adapt the results from the literature to the particular context of our experiment. In the third explanation, *biased experts*, the forecasts are in error because the experts themselves are biased. This bias could be due to the experts not providing their full effort or failing to rely on, or not knowing, the literature.

29. In DellaVigna and Pope (forthcoming) we consider further characteristics, such as citations and academic rank.

30. Our results are unlikely to be biased due to an atypical statistical draw, given the large sample size. We can quantify the magnitude of the sample error in the data by performing a Bayesian shrinkage correction (*e.g.* Jacob and Lefgren, 2008). For each treatment k we calculate $\hat{e}_{Shrink} = \frac{\bar{\sigma}^2}{\bar{\sigma}^2+\sigma_k^2}\hat{e}_k + \left(1 - \frac{\bar{\sigma}^2}{\bar{\sigma}^2+\sigma_k^2}\right)\bar{e}$, where $\bar{\sigma}^2$ is the variance across the eighteen effort estimates ($\hat{e}_k$) and σ_k^2 is the square of the estimated standard error of effort for treatment k. The estimator takes a convex combination between the estimated $\hat{e}_k$ (Table 3) and the average effort across all eighteen treatments ($\bar{e}$). As Online Appendix Figure 6 shows, this correction barely affects the point estimates, given that the standard errors for each treatment are small relative to the cross-treatment differences.

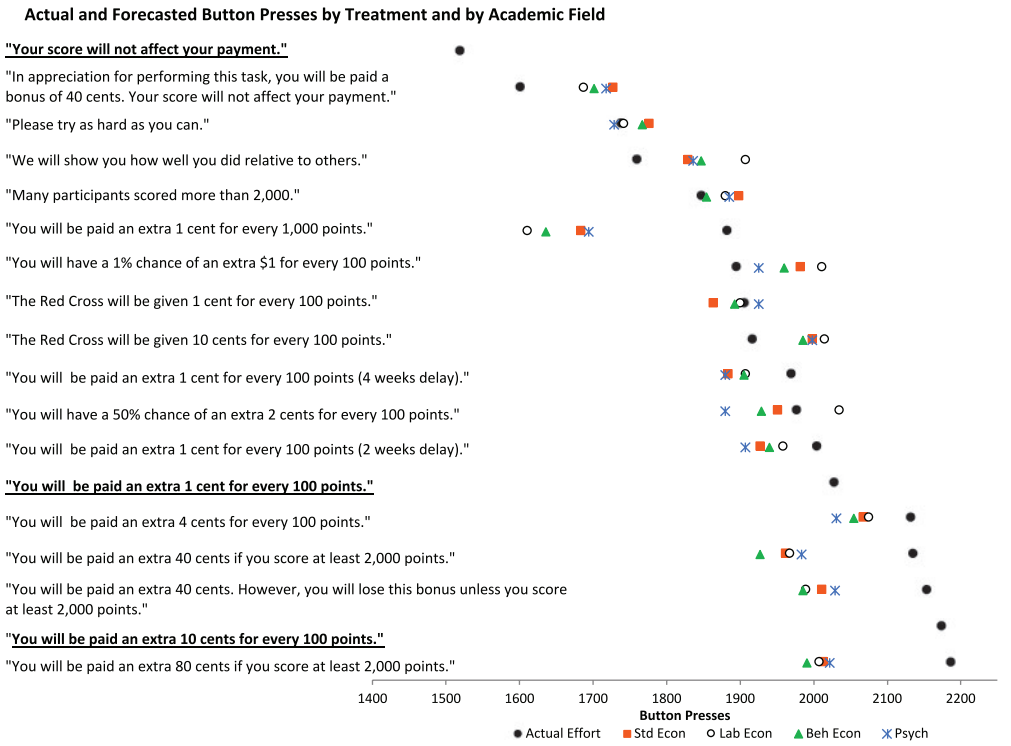

FIGURE 7

Average button presses by treatment and average expert forecasts, by academic field of expert

Notes: Figure 7 follows the same format of Figure 5, except that it splits the forecasts by the primary field of the 208 academic experts: behavioural economics, standard economics (consisting of applied microeconomics and economic theory), laboratory experiments, and psychology (which includes experts in behavioural decision-making).

To carefully discuss the three possible explanations above, we undertake a meta-analysis of related papers. We require: (1) a laboratory or field experiment (or natural experiment); (2) a treatment comparison that matches the one in our study; (3) an outcome variable about (broadly conceived) effort, such as responding to a survey.

The resulting data set includes 42 papers covering 8 of the 15 treatment comparisons, with the summary measures in Table 4 and the detailed paper-by-paper summaries in Online Appendix Table 2. The meta-analysis covers the treatments with very low pay (6 papers), charitable giving (5 papers), gift exchange (11 papers), probability weighting (4 papers), social comparisons a la Cialdini (9 papers), ranking (5 papers), and task significance (5 papers).

For each paper, we compute the treatment effect in standard deviation units (*i.e.* Cohen's d), with its standard error. We then generate the average Cohen's d across the papers using inverse-variance weighting, which is consistent with the fixed effect estimator commonly used in meta-analysis studies (Column 8 in Table 4). We also report an alternative Cohen's d weighting papers by their Google Scholar citations to capture the impact of prominent papers (Column 9). The table also reports the number of papers for a treatment (Column 5), and the number of papers with MTurk subjects or a similar online sample (Column 6). For comparison, the table also reports the treatment effects from our MTurk sample in standard deviation units (Column 3) as well as the average forecast in standard deviation units (Column 4).

TABLE 4
Experimental findings compared to meta-analysis of findings in literature

		Our results	Expert forecasts	Meta-analysis of literature (papers with similar treatments on effort)				
Category	Comparison	S.D. units (Cohen's d)	S.D. units (Cohen's d)	Number of papers	Papers with Mturk	Total sample size	Meta-analysis Cohen's d	Citation-weighted Cohen's d
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Very low pay	Compare very-low-pay (1c per 1,000 points) to no piece rate	0.521 (0.063)	0.196 (0.061)	6	0	1306	−0.059 (0.056)	−0.445 (0.170)
Social preferences: charity	Compare low piece rate to charity (1c) to low piece rate to self (1c)	−0.190 (0.060)	−0.211 (0.060)	5	0	1638	−0.076 (0.050)	0.026 (0.072)
	Compare high piece rate to charity (10c) to high piece rate to self (10c)	−0.434 (0.061)	−0.300 (0.061)	5	0	1574	−0.260 (0.051)	−0.263 (0.070)
	Compare high piece rate to charity (10c) to low piece rate to charity (1c)	0.018 (0.060)	0.166 (0.060)	5	0	1668	0.003 (0.049)	0.005 (0.068)
Social preferences: gift exchange	Compare gift exchange (40c) to no piece rate	0.114 (0.061)	0.265 (0.061)	11	0	3211	0.174 (0.041)	0.816 (0.243)
Probability weighting	Compare probabilistic piece rate (1% of $1) to deterministic piece rate with expected value (1c)	−0.202 (0.060)	−0.094 (0.060)	4	0	2355	−0.091 (0.042)	0.110 (0.099)
Social comparisons	Compare Cialdini-type comparison to no piece rate	0.447 (0.063)	0.487 (0.063)	9	0	243423	0.018 (0.005)	0.119 (0.034)
Ranking	Compare expectation of rank to no piece rate	0.334 (0.062)	0.457 (0.062)	5	0	1758	−0.032 (0.052)	0.232 (0.093)
Task significance	Compare task significance to no piece rate	0.314 (0.062)	0.337 (0.061)	5	2	1889	0.188 (0.047)	0.797 (0.176)

Notes: The table lists the eight treatments considered for our meta-analysis. Column (2) describes the treatment comparison for the control and treatment groups. For example, for the very-low-pay treatment, we compare a treatment with very low piece rate to a treatment with no piece rate. All treatment effect comparisons refer to comparing the two treatments, Columns (3) and (4) report the results of the experiment and expert forecast respectively, in units of Cohen's d (which we use as the standardized measure of effect size). Columns (5) through (7) report the summary statistics for our meta-analysis of each treatment, listing the total number of papers, the total number of papers with online workers and the total sample size for each treatment. The aggregate Cohen's d for our meta-analysis of each treatment in columns (8) and (9) are weighted averages across studies, where the weights used are the inverse-variance and Google Scholar citations respectively. For the Charity treatments, notice that one of the three comparisons is redundant with the others, since the set of papers is the same, but we report all three for clarity.

We stress two main caveats. First, despite our best efforts to track down papers, including contacting the authors of key papers for suggestions, it is sometimes difficult to determine whether a paper belongs to a treatment comparison and it is likely that we are missing some relevant papers. Second, the meta-analysis does not represent all treatments. It does not cover the 4-cent piece rate treatment since it is not a behavioural treatment and we already have a model-based benchmark. It also does not cover the gain–loss treatments because the forecast errors for those treatments are related to misforecasting the effect of a payoff threshold, not to poor forecasts of loss aversion.

Finally, we could not find any paper that considers how effort varies when the pay is immediate, versus delayed by about 2 weeks and 4 weeks.[31]

We highlight three features of this data set. First, we found only two papers using an online sample like MTurk; thus, the experts could not rely on experiments with a comparable sample. Second, nearly all papers contain only one type of treatment; papers such as ours and Bertrand *et al*. (2010) comparing a number of behavioural interventions are uncommon. Third, for most treatments we found only a few papers, sometimes little-known studies outside economics, including for classical topics such as probability weighting. Thus, an expert who wanted to consult the literature could not simply look up one or two familiar papers.

Turning to the meta-analysis, in the very-low-pay literature we find six papers, including Gneezy and Rustichini (2000), with both a very-low-piece-rate treatment and a no-piece-rate treatment. Some of the papers mention crowd out (such as Gneezy and Rey-Biel, 2014), while others do not, but in the context the pay is very low (*e.g*. Ashraf *et al*., 2014). The findings are split, with some papers finding a decrease in effort with very low pay, while other papers (like us) find a sizable increase in effort instead. The meta-analysis Cohen's d is slightly negative (−0.06 s.d.) and clearly negative if weighting by citations (−0.44 s.d.).

In the charitable giving literature, we consider papers comparing a piece rate to self versus the same piece rate for the charity, and also comparing a low piece rate to charity and a high piece rate to charity. Based on five papers with these features, we draw three comparisons: (1) piece rate to self versus to charity (low piece rate); (2) piece rate to self versus to charity (high piece rate); (3) low- versus high-piece rate to charity. The results in the first two comparisons vary sizably across the papers, but the latter comparison yields consistent results: there is generally no effort increase from increasing the return to the charity.

The gift exchange comparison has the largest number of papers we found (11 papers). The Cohen's d indicates a small, positive effect of 0.17 SD in response to a monetary gift. The effect is much larger when citation weighted, given the large effects in Gneezy and List (2006).

Next, we compare treatments with a probabilistic incentive (with low probability) to a certain incentive with the same expected value. Surprisingly, we found no papers in economics, but we located four papers on survey and test completion. The meta-analysis Cohen's d is −0.09.

For the social comparison, we draw on the meta-analysis in Coffman *et al*. (2016), and estimate a small, though statistically significant, Cohen's d of 0.02. This literature has by far the most precise Cohen's d estimates, given the large sample sizes.

Next, we consider experiments in which subjects are told that they will be ranked relative to others, with no incentive tied to the rank. These treatments on average yield no effect (Cohen's d of −0.03). By comparison, the task significance treatments yield a positive Cohen's d of 0.19 s.d., and a very large Cohen's d of 0.80 s.d. in the citation-weighted measure.

In Figure 8 we display, for each of the eight treatments, the average expert forecast and the effort implied by the meta-analysis, and relate these predictions to the actual results. Using this figure, one can identify several interesting cases that shed light on the three classes of explanations: *biased literature*, *biased context*, and *biased experts*.

Some treatments show reasonably accurate predictions by both experts and the literature (*e.g*. gift exchange). A different case is when the literature-based predictions are poor, but the experts are accurate (*e.g*. social comparison). In this case, the literature might be biased and the experts know it is biased and do not rely on it. Alternatively, the literature may be accurate but our context is different than the typical paper in the literature and the experts are able to adapt their knowledge from the literature to our new context.

31. Kaur *et al*. (2015) fits in the category, but their maximum distance to pay (payday) is 6 days. Designs such as Augenblick *et al*. (2015) vary the distance between the effort decision, and the effort itselt, not the distance to pay.

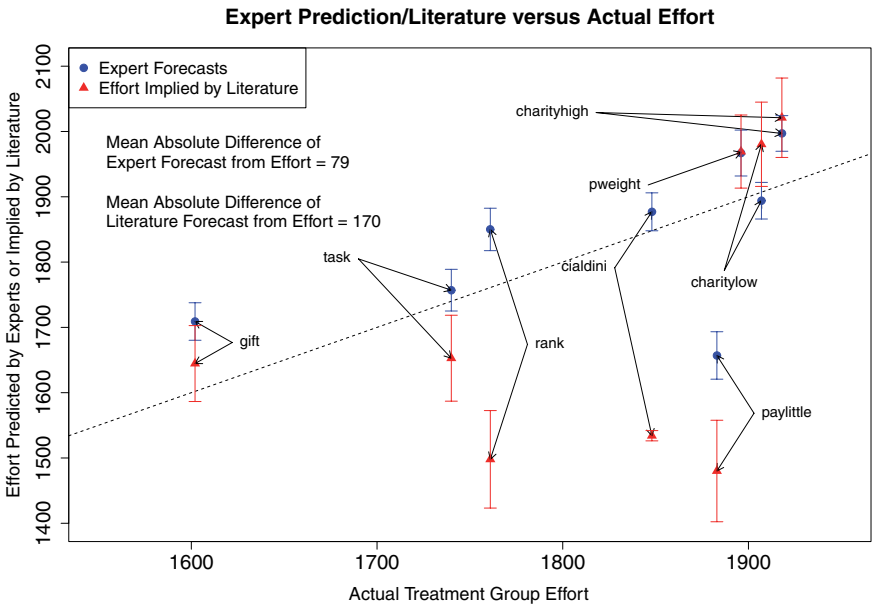

FIGURE 8

Prediction based on literature meta-analysis versus expert forecasts

Notes: Figure 8 presents a scatterplot of the eight treatments for which we conducted a meta-analysis, with the effort of the MTurk treatment group on the x-axis, and either the expert forecast or the effort implied by the literature on the y-axis. The literature-implied effort for a given treatment is the sum of the MTurk control group effort and the scaled aggregate Cohen's d in the literature (the latter being scaled by the pooled standard deviation of the efforts in the MTurk control and treatment groups). Error bars indicating 95% confidence intervals are plotted for the expert and literature forecasts. The figure also displays the 45° line, so the vertical distances between the points and this line represent the deviations of the expert or literature forecasts from the actual effort.

Another interesting case is when the literature makes accurate predictions while the experts are in error. When comparing effort with low and high return to charity, the literature (and our experimental results) finds no difference between the two treatments. Yet, the experts predict a 0.15 standard deviation higher effort with the high return to charity. In this case, the experts may be biased and fail to use the (recent) literature when making forecasts.

The final case is when the predictions by both the experts and the literature are inaccurate. In the very-low-pay condition, both experts and (especially) the literature under-predict the effort. This could be the result of a biased literature. Alternatively, the literature may be unbiased, but our context may be unique and the experts are unable to see that it will produce a result that is different than the one in the literature.

With only eight treatments, it is difficult to make a definitive claim about what is the most likely explanation for the differences between the expert forecasts and our experimental results. Indeed, we find some evidence in favour of each of the three classes of explanations.

An interesting comparison across the eight treatments is between the experts and the meta-analysis: do the experts outperform the forecasts formed based on the literature? The average absolute deviation between predictions and results is more than twice as large for the literature-based predictions than for the expert forecasts. This difference gets larger if the meta-analysis weighs papers based on their citation count (Online Appendix Figure 7). This puts further in perspective the quality of expert forecasts.

In principle, we would like to also relate the strength and precision of the evidence in the meta-analysis to the uncertainty in the expert forecasts. However, there is limited variation in the

strength of evidence as all but two of the treatment comparisons include four to six papers, with only gift exchange and social comparisons having twice as many. The precision in the Cohen's d estimate is also quite parallel across treatments, with the standard error equal to 0.04–0.05 SD for all treatments, other than for the social comparison treatments.

Using all fifteen treatments, we relate instead the heterogeneity in the expert forecasts to the heterogeneity of MTurker effort in that treatment. If the dispersion of forecasts among experts in a particular treatment reflects behavioural forces affecting effort in opposite directions, such as overweighting of small probabilities versus curvature of the utility function in the probabilistic pay treatment, and the contrasting behavioural forces differ across workers, treatments with high heterogeneity in forecasts may also have high heterogeneity in MTurker effort. Online Appendix Figure 8 provides evidence of a positive correlation among the fifteen treatments.[32]

As we discussed, the meta-analysis is limited to papers on effort, since we cannot directly translate evidence on other outcomes. Of course, if we had estimates of the underlying behavioural parameters in the literature, we could translate the estimate in effort units, given our estimates for the curvature of the cost of effort function and for the motivation term.

While we cannot do this for all treatments, we present estimates for the probabilistic piece rate treatments based on such structural estimates. In Online Appendix Table 3 we list key estimates of the probability weighting function (mostly from lottery choice), and derive the implied weights for 1% and 50% probabilities. Averaging across the papers, the probability weight for a 1% probability is 6%, while the probability weight for a 50% probability is 45%. Given our estimates for the cost function (Table 5, Column (1)) and assuming risk-neutrality, these values imply a predicted effort of 2,142 points in the 1% treatment and 2,022 points in the 50% treatment.[33] The latter estimate is close to the MTurk effort, possibly explaining why the experts guess accurately this treatment. The meta-analysis-based estimate for the 1% treatment is instead high relative to the data, plausibly contributing to the expert overestimation of the impact of this treatment.

Overall, in terms of explaining why the expert forecasts at times differ from our experimental results, we find pieces of evidence supporting each of the explanations—*biased literature*, *biased context*, and *biased experts*. Going forward, how do we gain a better understanding of why expert and literature-based forecasts may be biased? One option is to explore expertise more broadly as we do in the companion paper (DellaVigna and Pope, forthcoming), where we compare the forecasts made by experts to forecasts made by non-experts (undergraduate students, MTurk participants, etc.). This can help provide evidence on the treatments for which knowing the literature might lead to bias. In DellaVigna and Pope (forthcoming), we also look at different types of expertise, for example comparing experts who are familiar with the MTurk environment to those that are not. We find no evidence of a difference in forecast ability across these two groups, which is evidence against a *biased context* account.

Additional future work can try to further understand inaccuracies in expert forecasts. For example, one could study the same treatments as in our experiment but with a different task. If the treatments where experts had a large amount of forecast error in this article are the ones where the experimental results change significantly (in the direction of the forecasts), this would be evidence of biased context for the current paper. If the treatment effects are largely the same, then this suggests that biased experts or biased literature is the more likely story. Hopefully future research can continue to tease apart these various explanations for how good experts are at making forecasts and why they sometimes make poor predictions.

32. The correlation is muted if one restricts attention to the eight treatments in the meta-analysis.
33. The results for a utility function with curvature of 0.88 or even 0.7 are similar.

TABLE 5
Estimates of behavioural parameters I: Mturkers actual effort and expert beliefs

Cost of effort specification	Power cost of effort		Exponential cost of effort	
Estimation method	Minimum distance estimator on average effort	NLS on individual effort	Minimum distance estimator on average effort	NLS on individual effort
	(1)	(2)	(3)	(4)
Panel A. Estimate of model on effort in three benchmark treatments				
Curvature γ of cost	33.137	20.546	0.0158	0.0156
of effort function	(10.251)	(5.266)	(0.0054)	(0.0039)
Level k of cost	2.56E-112	5.12E-70	1.27E-16	1.71E-16
of effort function	(1.11E-69)	(2.36E-68)	(4.87E-12)	(14.1E-16)
Intrinsic motivation s	7.13E-07	3.17E-06	3.32E-06	3.72E-06
(cent per point)	(9.63E-06)	(8.16E-06)	(2.32E-05)	(8.81E-06)
N	1664	1664	1664	1664
Implied effort, 4-cent treatment		7.608		
(actual effort 2,132, log 7.602)	2116	(expected log effort)	2117	2115
Implied effort, low-pay treatment		7.429		
(actual effort 1,883, log 7.424)	1893	(expected log effort)	1883	1885 / 1881 / 1878

Panel B. Estimates of social preferences and time preferences

Cost of effort specification:	Power cost of effort			Exponential cost of effort			
Estimation method:	Minimum distance est.		NLS estimates	Minimum distance est.		NLS estimates	
	Estimate from Mturk (95% CI)	Median forecast (25th, 75th ptile)	Estimate from Mturk (95% CI)	Estimate from Mturk (95% CI)	Median forecast (25th, 75th ptile)	Estimate from Mturk (95% CI)	Median forecast (25th, 75th ptile)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Social preferences parameters							
Pure altruism coefficient α	0.003	0.068	0.006	0.003	0.067	0.004	0.070
	(−0.024,0.042)	(0.002,0.51)	(−0.024,0.037)	(−0.024,0.041)	(0.002,0.53)	(−0.018,0.025)	(0.002,0.53)
Warm glow coefficient a	0.125	0.022	0.182	0.143	0.031	0.143	0.035
(scale of 1 cent altruism)	(0.003,0.6)	(−8e-04,0.76)	(−0.176,0.54)	(0.005,0.61)	(−1E-4,0.71)	(−0.132,0.417)	(4.5e-05,0.74)
Gift exchange Δs	3.26E-06	1.30E-05	2.04E-05	8.59E-06	2.80E-05	2.35E-05	3.00E-05
(cent per point)	(−7E-8,8E-5)	(2E-6,2E-4)	(−7E-5,1E-4)	(−2E-7,1.3E-4)	(4.5E-6,2.8E-4)	(−7E-5,1.2E-4)	(5E-6,3E-4)
Time preference parameters							
Present bias β	1.17	0.76	1.36	1.15	0.76	1.24	0.79
	(0.09,11)	(0.27,1.2)	(−1.56,4.28)	(0.09,9.9)	(0.29,1.2)	(−1.3,3.77)	(0.3,1.2)
(Weekly) discount factor δ	0.75	0.85	0.75	0.76	0.85	0.75	0.86
	(0.33,1.5)	(0.62,1)	(0.261,1.24)	(0.34,1.4)	(0.65,1)	(0.283,1.22)	(0.65,1)

Notes: Panel A reports the structural estimates of the model in Section 2. Columns (1) and (3) use a minimum-distance estimator employing three moments (average effort in three benchmark treatments) and three parameters, and is thus exactly identified. We estimate the model under two assumptions, a power cost of effort function (Column (1)) and an exponential cost of effort function (Column (3)). The standard errors are derived via a bootstrap with 1,000 draws. Columns (2) and (4) use a NLS specification using the individual effort of MTurkers (rounded to the nearest 100) in the three benchmark treatments. In the rows displaying the implied effort we compute the predicted effort given the parameters for the 4-cent treatment and the low-pay treatment. For the low-pay treatment in Column 4, in addition we present two alternative predictions which explicitly model the discontinuity in payoffs, with very similar results (see Appendix A for details). Panel B presents estimates for the behavioural parameters. Columns (1) and (4) use cost function parameters (from panel A columns (1) and (3) respectively), and relevant treatments moments to calculate the behavioural parameters. The confidence intervals are derived by bootstrapping the workers and redoing the same exercise, including estimation of the cost function parameters, then taking the 2.5th and 97.5th percentiles. Columns (3) and (6) are the results of an NLS estimation using all relevant moments (benchmark plus gift exchange, time-delay and charitable giving). Columns (2), (5), and (7) calculate the implied behavioural parameters using the cost function parameters from panel A columns, and the individual expert's forecast per each relevant treatment — we report the median, 25th, and 75th percentile implied parameters. We do not elicit parameters for the experts under the power cost function for the NLS estimate since we did not ask for the expected log effort, which is the key variable for that model.

7. ESTIMATES OF BEHAVIOURAL PARAMETERS

An advantage of field experiments is that their design can be tailored to a model, so as to test the model and estimate parameters. Surprisingly, model-based field experiments are still relatively

uncommon (Card *et al*., 2011). One of the difficulties of conducting these experiments is that the researcher needs to estimate a set of nuisance parameters (*e.g*. about the environment), to focus on the parameters of interest.

In our setting, the simplicity of the chosen task implies that the only nuisance parameters are those on the cost of effort. We thus designed the piece rate treatments to pin down these parameters, as stressed in Section 4. Armed with these estimates, we can identify the behavioural parameters of interest. Furthermore, since we informed the experts about the results in the benchmark treatments, we can, at least in principle, assume that the forecasters approximately share the estimates for these nuisance parameters. We now present the estimation procedures, and the resulting estimates, with additional details in Online Appendix A.

7.1. *Minimum-distance estimation*

For the minimum-distance estimation, we use as moments the average effort in the three benchmark treatments (no-pay, 1-cent, and 10-cent) to estimate $\hat{\gamma}$, $\hat{s}$, and $\hat{k}$. Panel A of Table 5 presents the estimates with power cost (Column 1) and exponential cost (Column 3), as we discussed in Section 4.

Given these estimates, we then back out the behavioural parameters using the average effort in the relevant behavioural treatments as moments. For example, assuming a power cost function, effort in the 1-cent and 10-cent charitable giving treatments equal

$$\bar{e}_{CH.01} = \left(\frac{\hat{s} + (\hat{a} + \hat{\alpha}) * .01}{\hat{k}} \right)^{1/\hat{\gamma}} \text{ and } \bar{e}_{CH.10} = \left(\frac{\hat{s} + \hat{a} * .01 + \hat{\alpha} * .10}{\hat{k}} \right)^{1/\hat{\gamma}}. \tag{11}$$

The system of two equations in two unknowns (given the estimates of $\hat{\gamma}$, $\hat{s}$, and $\hat{k}$) yields solutions for $\hat{\alpha}$ and $\hat{a}$. By design, the model is just identified. We derive confidence intervals for the parameters using a bootstrap procedure.

The appeal of this simple identification strategy is that the forecasters could also, at least in principle, have obtained the same estimates for $\hat{\gamma}$, $\hat{s}$, and $\hat{k}$, given the observed effort in the benchmark treatments. Under this assumption, we can take the forecasts $(e^i_{CH.01}, e^i_{CH.10})$ of expert i and back out the implied beliefs about social preferences $(\widetilde{\alpha}_i, \widetilde{a}_i)$ of expert i.

7.2. *Non-linear least squares*

The minimum-distance estimates assume no error term and thus, counterfactually, no heterogeneity in effort. It also assumes, for simplicity, that the incentives accrue continuously, as opposed to at fixed 100-point intervals. We now relax these assumptions using data on the individual-level effort.

We allow for a heterogeneous marginal cost of effort $c(e)$ in maximization problem (1). Namely, for the power cost case we assume that worker j has $c_j(e_j) = k e_j^{1+\gamma} (1+\gamma)^{-1} \exp(-\gamma \varepsilon_j)$, with ε_j normally distributed $\epsilon_j \sim N(0, \sigma_\epsilon^2)$. The additional noise term $\exp(-\gamma \epsilon_j)$ has a lognormal distribution, ensuring positive realizations for the marginal cost of effort. As DellaVigna *et al*. (2015) show, this implies the first-order condition $s + p - k e_j^\gamma \exp(-\gamma \epsilon_j) = 0$ and, after taking logs and transforming,

$$\log(e_j) = \frac{1}{\gamma} [\log(s+p) - \log(k)] + \epsilon_j. \tag{12}$$

Equation (12) can be estimated with non-linear least squares (NLS). Similarly, for the case of exponential cost function we assume $c(e_j) = k \exp(\gamma e_j) \gamma^{-1} \exp(-\gamma \varepsilon_j)$, yielding a parallel

estimating expression but with effort, rather than log effort, as dependent variable:

$$e_j = \frac{1}{\gamma}[\log(s+p) - \log(k)] + \epsilon_j. \tag{13}$$

The NLS estimation allows us to model the heterogeneity in effort e_j. To take into account the discontinuous incentives, we assume that the individual chooses output in units of 100 points, and estimate the model using output rounded to the closest 100-point: that is, a score of 2,130 points is recorded as 21 units of 100 points. This assumption allows us to use the first-order condition for effort and thus the NLS for estimation.[34]

Columns 2 and 4 of Panel A in Table 5 display the estimates of the NLS model using the benchmark treatments. The parameter estimates for the exponential cost function case (Column 4) are nearly identical to the minimum-distance ones (Column 3). The model perfectly fits the benchmark treatments and makes predictions for the 4-cent treatment and for the low-pay treatment that are very similar to the minimum distance ones.[35]

The NLS estimates for the power cost function (Column 2) yield a lower curvature than the minimum-distance estimates ($\hat{\gamma}_{NLS} = 20$ versus $\hat{\gamma}_{MD} = 33$). The NLS model, as (12) stresses, matches the expected log effort, while the minimum-distance matches the log of expected effort (given the assumed homogeneity). Nonetheless, both models fit the in-sample moments perfectly and make similar predictions for the 4-cent treatment and the low-pay treatment.[36]

We use the NLS estimator to estimate the behavioural parameters in Panel B. Formally, we run a NLS regression including the benchmark treatments as well as the behavioural treatments. We report the point estimates for the behavioural coefficients (Columns 3 and 6) and, for the exponential case, the behavioural parameters implied by the expert forecasts (Column 7).[37]

7.3. *Social preferences*

Returning to social preferences, equation (11) clarifies the difference between our models of altruism and warm glow: the altruism parameter α multiplies the actual return to the charity while the warm glow term a multiplies a constant return which we set, for convenience, to 0.01, the 1-cent return. Taking logs of output and differencing, we obtain

$$\log(\bar{e}_{CH.10}) - \log(\bar{e}_{CH.01}) = \frac{1}{\hat{\gamma}}\left[\log\left(\hat{s} + \hat{a} * .01 + \hat{\alpha} * .10\right) - \log\left(\hat{s} + \left(\hat{a} + \hat{\alpha}\right) * .01\right)\right].$$

The increase in output between the two treatments identifies the altruism parameter α, since the two right-hand side log terms differ only in the terms $\hat{\alpha} * 0.10$ versus $\hat{\alpha} * 0.01$. The warm glow parameter $\hat{a}$ is identified from the level of effort in the 1-cent charity treatment. The expression also clarifies that $1/\hat{\gamma}$ is the elasticity of effort with respect to motivation.

The altruism coefficient from the MTurk effort is estimated to be essentially zero in all four specifications, for example $\hat{\alpha} = 0.003$ in Column 1. Importantly, the confidence interval is tight

34. This is still an approximation, given that the choice of units is still discrete so strictly speaking the first-order condition does not apply.

35. The implied effort for the low-pay treatment still assumes an incentive of 0.1 cent every 100 point, rather than an incentive ocurring only every 1,000 points. In Appendix A we show that modelling the discrete jumps at 1,000 gives similar results for the implied effort in the low-pay treament.

36. Notice that for the NLS model with power cost in Column 2 of Table 5, the predictions are evaluated using the average log effort.

37. For the power cost case we cannot infer the parameters implied by the expert forecasts since we did not elicit the expected log points, as the model requires.

enough that we can reject even small values, such as the workers putting 0.04 as much weight on the charity as on themselves (Column 1). Instead, the median expert expects altruism $\widetilde{\alpha}_{med} = 0.068$ (Columns 2 and 5), outside the confidence interval of the MTurk estimates.

The pattern for warm glow is the converse: the worker effort indicates sizable warm glow, with a weight $\hat{a}$ between 0.12 (Column 1) and 0.18 (Column 3) on the average return for the charity. The median forecast instead is $\widetilde{a}_{med} = 0.02$ (Column 2), which is barely inside the 95% Confidence interval for the estimates from the MTurk effort. Online Appendix Figures 9 a and b show the distribution of the social preferences parameters $(\widetilde{\alpha}_i, \widetilde{a}_i)$ estimated from the 208 expert forecasts from the minimum-distance power cost specification (Column 1). The green solid line denotes the value implied by the median forecast, and the red dashed line indicates the parameter value implied by the actual MTurk worker effort.

Panel B of Table 5 also reports the estimated shift in motivation due to gift exchange. The impact on motivation is estimated to be tiny, consistent with the small gift exchange effect, as well as the small value for baseline motivation. We do not report the other motivation shift parameters in response to the other non-monetary treatments, but the estimates are similarly small in magnitude. The expert forecasts are generally in line, though some experts expect a sizeable shift in motivation due to the treatments.

7.4. *Time preferences*

We model effort in the delayed-payment treatments as in (7), with t denoting the weeks of delay, β the present bias parameter, and δ the (weekly) discount factor. As Panel B of Table 5 indicates, the estimates of the time preference parameters from the worker effort are noisy: the point estimate indicates no present bias, but the confidence intervals for β are wide.[38] Even given the imprecise estimate from the MTurk data, there is useful information in the expert forecasts: the median expert (Column 2) expects present bias ($\widetilde{\beta}_{med} = 0.76$) with a significant left tail of smaller estimates (as well as estimates above 1).

7.5. *Probability weighting*

In prospect theory, the probability weighting function $\pi(P)$ transforms probabilities P into weights, which are then used to calculate the value of the "prospects". The evidence on probability weighting (*e.g.* Prelec, 1998, and see Online Appendix Table 3) suggests that small probabilities are overweighted by a factor of 3–6, with a probability of 50% slightly downweighted. The treatment with a 1% probability of a \$1 piece rate allows us to test for such overweighting of small probability and estimate $\pi(0.01)$. The design also includes a treatment with 50% probability of a 2-cent piece rate to provide evidence on the concavity of the value function, that is, the risk aversion.

We model optimal effort in the probabilistic treatments as in (10), allowing for a possibly concave utility function $u(p) = p^{\theta}$. This includes linear utility ($\theta = 1$), assumed so far, as well as the calibrated value $\theta = 0.88$ from Tversky and Kahneman (1991). For simplicity, we assume that the probability weight does not transform the 50% probability ($\pi(0.5) = 0.5$).

Since allowing for curvature in the utility function $u(p)$ affects the estimates also in the benchmark treatments, we re-estimate also the baseline parameters using the three benchmark treatments and the two probabilistic treatments. In Table 6, Panel A, we report the results for the NLS estimates; the results are similar with minimum distance.

38. The lack of support for present bias may also reflect the 24-hour delay in pay (Balakrishnan *et al*., 2016).

TABLE 6
Estimates of reference-dependent parameters: Mturker actual effort and expert beliefs

Estimation method	NLS on individual effort in three treatments					
Cost of effort specification	Power cost of effort			Exponential cost of effort		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Estimate of model on effort in three benchmark treatments and two probability treatments						
Curvature γ of cost	20.59	18.87	19.64	0.0134	0.0119	0.0072
of effort function	(4.22)	(3.92)	(14.19)	(0.0024)	(0.0021)	(0.0027)
Level k of cost	3.77E-70	3.92E-64	1.02E-66	2.42E-14	7.50E-13	5.46E-08
of effort function	(1.45E-68)	(1.16E-62)	(1.12E-64)	(1.19E-13)	(3.27E-12)	(3.50E-7)
Intrinsic motivation s	2.66E-06	6.22E-06	3.75E-6	1.65E-5	5.55E-5	3.14E-3
(cent per point)	(5.45E-6)	(12.24E-06)	(34.61E-6)	(2.26E-5)	(6.78E-5)	(7.16E-3)
Probability weighting π (1%)	0.19%	0.38%	0.30%	0.24%	0.47%	4.30%
(in %)	(0.15)	(0.26)	(1.31)	(0.14)	(0.24)	(5.25)
Curvature of utility over	1.00	0.88	0.92	1.00	0.88	0.47
piece rate	(assumed)	(assumed)	(0.79)	(assumed)	(assumed)	(0.23)
N	2,787	2,787	2,787	2,787	2,787	2,787
Implied probability weighting $\pi(1\%)$ by experts						
25th percentile	.	.	.	0.05%	0.11%	1.7%
Median	.	.	.	1.5%	2.3%	12%
75th percentile	.	.	.	5.4%	7.6%	24%

Panel B. Estimate of loss aversion based on local approximation	Estimate from Mturk (95% CI)	Median Forecast (25th, 75th ptile)
	(1)	(2)
Reference dependence parameter		
loss aversion λ	1.73	2.76
	(0.2,4.9)	(2,4.4)

Notes: Panel A reports the structural estimates of the model in Section 2 using a NLS regression for observations in the three benchmark treatments and in the two probabilistic pay treatments. We estimate the model under two assumptions, a power cost of effort function (Columns 1–3) and an exponential cost of effort function (Columns 4–6). The specification reports the estimate for a probability weighting coefficient under the assumption of linear value function (Columns 1 and 4), concave value function with curvature 0.88 as in Tversky and Kahneman (Columns 2 and 5) and with estimated curvature (Columns 3 and 6). Panel B shows the estimates for the loss aversion parameter, which is obtained with a local approximation, see text.

The probability weight for a 1% probability is estimated to be *smaller* than 1% under the assumption of either linear utility (Columns 1 and 4) or concave utility with the Kahneman and Tversky curvature (Columns 2 and 5). Thus, we do not find evidence of overweighting of small probabilities. In contrast, the median expert expects overweighting of 1 percent probability under either specification (Columns 4 and 5). The difference between the median forecast and the estimate from the MTurk effort is statistically significant.

The specification with estimated curvature of the utility function (Columns 3 and 6) leads to imprecise results, yielding very high curvature with the exponential cost function (Column 6) and near-linear utility with power cost function (Column 3). The former case, given the high curvature of the value function, is the only case with estimates implying overweighting of small probability, but the estimates are very imprecise.

Thus, under plausible curvature of the value function, the MTurk effort does not provide evidence of overweighting of small probabilities, contrary to the forecast of the median expert.

7.6. *Loss aversion*

We estimate the loss aversion parameter λ using the three gain–loss treatments. The experts are quite off in their forecasts of these treatments because it was difficult to predict the impact of a threshold payment at 2,000 points.[39] For the estimation, we derive an approximation that bypasses this misprediction. We compare the difference between the 40-cent loss treatment and the 40-cent gain treatment $e_{L.40} - e_{G.40}$, and the difference between the 80-cent gain treatment and the 40-cent gain treatment, $e_{G.80} - e_{G.40}$. As we show in Online Appendix A, the following approximation holds

$$\frac{e_{L.40} - e_{G.40}}{e_{G.80} - e_{G.40}} \simeq \frac{(\lambda - 1)\eta}{1 + \eta}.$$

Under the standard assumption of unitary gain utility ($\eta = 1$), this expression allows for estimation of the loss aversion λ.[40]

The distribution of the loss aversion parameter $\widetilde{\lambda}_i$ according to the experts is broadly centered around 2.5–3, with a median $\widetilde{\lambda}_{med} = 2.76$ (Table 6, Panel B). Thus, experts hold beliefs in line with the Tversky and Kahneman (1991) calibration which, revisited in the Koszegi and Rabin (2006) formulation, implies a loss aversion parameter of $\lambda = 3$ (assuming $\eta = 1$). The estimate from the MTurk worker effort is smaller, $\hat{\lambda} = 1.73$, but with a wide confidence interval including the value $\lambda = 3$. Unfortunately, the estimate for λ is quite noisy because the impact of going from the 40 cent gain treatment to the 80 cent gain treatment is quite small, making it hard to compare to the effect of the 40 cents loss treatment.

7.7. *Robustness*

In Online Appendix Table 4 we explore the robustness to alternative specifications, under the maintained NLS specification with exponential cost of effort function. We examine the impact of mis-specification in the cost function by forcing the curvature parameter γ to the values of 0.01 (Column 1) and 0.02 (Column 2). We also allow for curvature of the value function with concavity $\theta = 0.88$ when estimating the parameters (Column 3). We also use continuous points assuming that the piece rates are paid continuously (Column 4). These changes have limited impact on the estimates, other than on the coefficient β which is more sensitive, not surprisingly given the wide confidence intervals in the benchmark estimates.

8. CONCLUSION

What motivates workers in effortful tasks? How do different monetary and non-monetary motivators compare in effectiveness? Do the results line up with the expectations of researchers?

We present the results of a large-scale real-effort experiment on MTurk workers. The model-based eighteen arm experiment compares three classes of motivators: (1) standard incentives

39. In hindsight, we should have offered the results of the 40 cent gain treatment as a fourth benchmark.

40. Unlike the other derivations, this solution is an approximation. However, given that the differences in effort between the threshold treatments are small, the bias in estimate due to the approximation should be small as well. Given that the estimation is based on a ratio, we only use observations in which the denominator is positive and larger than 10 units of effort, since smaller differences may be hard for experts to even control with a mouse. We also do not include observations with negative λ.

in the form of piece rates; (2) behavioural factors like present bias, reference dependence, and social preferences, and (3) non-monetary inducements more directly borrowed from psychology.

Monetary incentives work as expected, including a very low piece rate which does not crowd out motivation. The evidence is partly consistent with behavioural models, including loss aversion and warm glow, but we do not find evidence of overweighting of small probabilities. The psychological motivators are effective, though less so than monetary incentives.

We then compare the results to forecasts by 208 behavioural experts. The experts on average anticipate several key features of the data, like the effectiveness of psychological motivators compared to the effectiveness of incentives. A sizeable share of the experts, however, expect crowd-out, probability weighting, and pure altruism, unlike what we observe in the data. Compared to the predictions one would have made based on a meta-analysis of related treatments in the literature, expert forecasts are more accurate and less noisy predictors of the results.

An important caveat is that the relative effectiveness of the various treatments may be context dependent. Some treatments that had a limited effect in our context, such as probabilistic piece rates, may have large effects in a different task or with a different participant pool. As always, it will be important to see replications. By estimating the behavioural parameters, we set up a methodology to compare effects across different settings and subject pools.

Further, while we have studied a large set of motivators, it is by no means an exhaustive list. For example, we did not include treatments related to limited attention and salience, left-digit bias, or self-affirmation. In addition, our focus has been on costly effort, but future work could consider other outcomes, like contributions to public goods. Future research should also investigate for what questions and policies experts are more likely to make accurate forecasts.

Finally, the combination of head-to-head comparisons of treatments and expert forecasts can help inform the role of behavioural economists in helping policy-makers or businesses. For example, one of the authors worked with a non-profit company that was trying to motivate its clients to refinance their homes. The company wanted advice on the design of a letter to maximize take up. But how informed is our advice? Should they follow it?

It would seem that an alternative to using forecasts is run an experiment randomizing alternative options. But even in a setting in which an organization can run randomized trials, it will only test a subset of treatments. Which treatments are chosen for randomization once again will depend on implicit (or explicit) forecasts of effectiveness. Thus, we expect that the study of horseraces of treatments, and of forecasts, is with us to stay.

Acknowledgments. We thank Ned Augenblick, Oriana Bandiera, Dan Benjamin, Jordi Blanes-i-Vidal, Patrick Dejarnette, Jon de Quidt, Clayton Featherstone, Judd Kessler, David Laibson, John List, Benjamin Lockwood, Barbara Mellers, Katie Milkman, Don Moore, Sendhil Mullainathan, Victoria Prowse, Jesse Shapiro, Uri Simonsohn, Erik Snowberg, Philipp Strack, Justin Sydnor, Dmitry Taubinsky, Richard Thaler, Mirco Tonin, and Kevin Volpp. We are also grateful to the audiences at Bonn University, Frankfurt University, the FTC, the London School of Economics, the Max Planck Institute in Bonn, the University of Toronto, at the University of California, Berkeley, the University of Santiago, Yale University, the Wharton School, at the 2016 JDM Preconference, the 2015 Munich Behavioural Economics Conference and at the 2016 EWEBE conference for useful comments. We also thank Alden Cheng, Thomas Graeber, Johannes Hermle, Jana Hofmeier, Lukas Kiessling, Tobias Raabe, Michael Sheldon, Avner Shlain, Jihong Song, Alex Steiny, Patricia Sun, and Brian Wheaton for excellent research assistance. We are also very thankful to all the experts who took the time to contribute their forecasts. We are very grateful for support from the Alfred P. Sloan Foundation (award FP061020).

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

ALLEN, E. J., DECHOW, P. M., POPE, D. G. *et al*. (2017), "Reference-dependent Preferences: Evidence from Marathon Runners", *Management Science*.
AMIR, O. and ARIELY, D. (2008), "Resting on Laurels: The Effects of Discrete Progress Markers as Subgoals on Task Performance and Preferences", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **34**, 1158–1171.
ANDREONI, J. (1989), "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence", *Journal of Political Economy*, **97**, 1447–1458.
ANDREONI, J. (1990), "Impure Altruism and Donations to Public Goods: A Theory of Warm-glow Giving", *The Economic Journal*, **100**, 464–477.
ANDREONI, J. and SPRENGER, C. (2012), "Estimating Time Preferences from Convex Budgets", *American Economic Review*, **102**, 3333–3356.
ASHRAF, N., BANDIERA, O. and JACK, B. K. (2014), "No Margin, No Mission? A Field Experiment on Incentives for Pro-Social Tasks", *Journal of Public Economics*, **120**, 1–17.
AUGENBLICK, N., NIEDERLE, M. and SPRENGER, C. (2015), "Working Over Time: Dynamic Inconsistency in Real Effort Tasks", *Quarterly Journal of Economics*, **130**, 1067–1115.
BALAKRISHNAN, U., HAUSHOFER, J. and JAKIELA, P. (2016), "How Soon Is Now? Evidence of Present Bias from Convex Time Budget Experiments" (IZA Discussion Paper #9653).
BANERJEE, A., CHASSANG, S. and SNOWBERG, E. (2016), "Decision Theoretic Approaches to Experiment Design and External Validity", *Handbook of Field Experiments*. Forthcoming.
BANDIERA, O., BARANKAY, I., RASUL, I. (2013), "Team Incentives: Evidence from a Firm Level Experiment", *Journal of the European Economic Association*, **11**, 1079–1114.
BARSEGHYAN, L., MOLINARI, F., O'DONOGHUE, T. *et al*. (2013), "The Nature of Risk Preferences: Evidence from Insurance Choices", *American Economic Review*, **103**, 2499–2529.
BECKER, G. S. (1974), "A Theory of Social Interactions", *Journal of Political Economy*, **82**, 1063–1093.
BERGER, J. and POPE, D. (2011), "Can Losing Lead to Winning", *Management Science*, **57**, 817–827.
BERTRAND, M., KARLAN, D., MULLAINATHAN, S. *et al*. (2010), "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment", *Quarterly Journal of Economics*, **125**, 263–306.
CAMERER, C., DREBER, A., FORSELL, E. *et al*. (2016), "Evaluating Replicability of Laboratory Experiments in Economics" *Science*, **351**, 1433–1436.
CARD, D., DELLAVIGNA, S. and MALMENDIER, U. (2011), "The Role of Theory in Field Experiment", *Journal of Economic Perspectives*, **25**, 39–62.
COFFMAN, L. and NIEHAUS, P. (2014), "Pathways of Persuasion" (Working Paper).
CONLIN, M., O'DONOGHUE, T. and VOGELSANG, T. J. (2007). "Projection Bias in Catalog Orders", *American Economic Review*, **97**, 1217–1249.
DECI, E. L. (1971), "Effects of Externally Mediated Rewards on Intrinsic Motivation", *Journal of Personality and Social Psychology*, **18**, 105–115.
DELLAVIGNA, S. (2009), "Psychology and Economics: Evidence from the Field", *Journal of Economic Literature*, **47**, 315–372.
DELLAVIGNA, S., LIST, J. A. and MALMENDIER, U. (2012), "Testing for Altruism and Social Pressure in Charitable Giving", *Quarterly Journal of Economics*, **127**, 1–56.
DELLAVIGNA, S., LIST, J., MALMENDIER, U. and RAO, G. (2015), "Estimating Social Preferences and Gift Exchange at Work" (Working Paper).
DELLAVIGNA, S. and POPE, D. (forthcoming), "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy*
DREBER, A., PFEIFFER, T., ALMENBERG, J. *et al*. (2015), "Using Prediction Markets to Estimate the Reproducibility of Scientific Research", *PNAS*, **112**, 15343–15347.
EREV, I., ERT, E., ROTH, A. E. *et al*. (2010), "A Choice Prediction Competition: Choices from Experience and from Description", *Journal of Behavioural Decision Making*, **23**, 15–47.
COFFMAN, L., FEATHERSTONE, C. R. and KESSLER, J. (2016), "A Model of Information Nudges" (Working Paper).
FEHR, E. and GÄCHTER, S. (2000), "Fairness and Retaliation - The Economics of Reciprocity", *Journal of Economic Perspectives*, **14**, 159–181.
FEHR, E., KIRCHSTEIGER, G. and RIEDL, A. (1993), "Does Fairness Prevent Market Clearing? An Experimental Investigation", *Quarterly Journal of Economics*, **108**, 437–459.
FRANK, R. H. (1985), *Choosing the Right Pond: Human Behavior and the Quest for Status* (New York: Oxford University Press).
FREDERICK, S., LOEWENSTEIN, G. and O'DONOGHUE T. (2002), "Time Discounting and Time Preference: A Critical Review", *Journal of Economic Literature*, **40**, 351–401.
FRYER, R. Jr., LEVITT, S. D., LIST, J. A. *et al*. (2012), "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment" (NBER Working Paper No. 18237).
GNEEZY, U. and LIST, J. A. (2006), "Putting Behavioural Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments" *Econometrica*, **74**, 1365–1384.
GNEEZY, U. and REY-BIEL, P. (2014), "On the Relative Efficiency of Performance Pay and Noncontingent Incentives", *Journal of the European Economic Association*, **12**, 62–72.

GNEEZY, U. and RUSTICHINI, A. (2000), "Pay Enough or Don't Pay at All", *Quarterly Journal of Economics*, **115**, 791–810.
GRANT, A. M. (2008), "The Significance of Task Significance: Job Performance Effects, Relational Mechanisms, and Boundary Conditions", *Journal of Applied Psychology*, **93**, 108–124.
GROH, M., KRISHNAN, N., MCKENZIE, D. *et al*. (2015), "The Impact of Soft Skill Training on Female Youth Employment: Evidence from a Randomized Experiment in Jordan" (Working Paper).
HORTON, J. J. and CHILTON, L. B. (2010), "The Labor Economics of Paid Crowdsourcing", *Proceedings of the 11th ACM Conference on Electronic Commerce*.
HORTON, J. J., RAND, D. and ZECKHAUSER, R. (2011), "The Online Laboratory: Conducting Experiments in a Real Labor Market", *Experimental Economics*, **14** 399–425.
HOSSAIN, T. and LIST, J. A. (2012), "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations", *Management Science*, **58**, 2151–2167.
IMAS, A. (2014). "Working for the "warm glow": On the Benefits and Limits of Prosocial Incentives", *Journal of Public Economics*, **114**, 14–18.
IPEIROTIS, P. G. (2010), "Analyzing the Amazon Mechanical Turk Marketplace", *XRDS: Crossroads, The ACM Magazine for Students*, **17**, 16–21.
JACOB, B. and LEFGREN, L. (2008), "Principals as Agents: Subjective Performance Assessment in Education", *Journal of Labor Economics*, **26**, 101–136.
KAHNEMAN, D. and TVERSKY, A. (1979), "Prospect Theory: An Analysis of Decision Under Risk", *Econometrica*, **47**, 263–292.
KAUR, S., KREMER, M. and MULLAINATHAN, S. (2015), "Self-Control at Work" *Journal of Political Economy*, **123**, 1227–1277
KOSZEGI, B. (2014), "Behavioural Contract Theory" *Journal of Economic Literature*, **52**, 1075–1118.
KOSZEGI, B. and RABIN, M. (2006), "A Model of Reference-Dependent Preferences", *Quarterly Journal of Economics*, **121**, 1133–1165.
KUZIEMKO, I., NORTON, M. I., SAEZ, E. *et al*. (2015), "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments", *American Economic Review*, **105**, 1478–1508.
LAIBSON, D. (1997), "Golden Eggs and Hyperbolic Discounting", *Quarterly Journal of Economics*, **112**, 443–477.
LAIBSON, D., REPETTO, A. and TOBACMAN, J. (2007), "Estimating Discount Functions with Consumption Choices over the Lifecycle" (Working Paper).
LOEWENSTEIN, G., BRENNAN, T. and VOLPP, K. G. (2007), "Asymmetric Paternalism to Improve Health Behaviors", *Journal of the American Medical Association*, **298**, 2415–2417.
MASLOW, A. H. (1943), "A Theory of Human Motivation", *Psychological Review*, **50**, 370–396.
MELLERS, B., STONE, E., MURRAY, T. *et al*. (2015), "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions", *Perspectives on Psychological Science*, **10**, 267–281.
O'DONOGHUE, E. and RABIN, M. (1999), "Doing It Now or Later", *American Economic Review*, **89**, 103–124.
PAOLACCI, G. 2010. "Running Experiments on Amazon Mechanical Turk", *Judgement and Decision Making*, **5**, 411–419.
PAOLACCI, G. and CHANDLER, J. (2014), "Inside the Turk: Understanding Mechanical Turk as a Participant Pool", *Current Directions in Psychological Science*, **23**, 184–188.
PRELEC, D. 1998. "The Probability Weighting Function", *Econometrica*, **66**, 497–527.
RABIN, M. (1998), "Psychology and Economics", *Journal of Economic Literature*, **36**, 11–46.
REES-JONES, A. (2014), "Loss Aversion Motivates Tax Sheltering: Evidence from US Tax Returns" (Working Paper).
ROSS, J., IRANI, L., SILBERMAN, M. S., *et al*. (2010), "Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk", in CHI '10 Extended Abstracts on Human Factors in Computing Systems, 2863–2872.
SANDERS, M., MITCHELL, F. and CHONAIRE, A. N. (2015), "Just Common Sense? How Well Do Experts and Lay-people Do at Predicting the Findings of Behavioural Science Experiments" (Working Paper).
SCHULTZ, P. W., NOLAN, J. M., CIALDINI, R. B. *et al*. (2007), "The Constructive, Destructive, and Reconstructive Power of Social Norms", *Psychological Science*, **18**.
SIMMONS, J. P. NELSON, L. D. and SIMONSOHN, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant", *Psychological Science*, **22**, 1359–1366.
SNOWBERG, E., WOLFERS, J. and ZITZEWITZ, E. (2007), "Partisan Impacts on the Economy: Evidence from Prediction Markets and Close Elections", *Quarterly Journal of Economics*, **122**, 807–829.
TETLOCK, P. E. and GARDNER, D. (2015), *Superforecasting: The Art and Science of Prediction* (New York: Crown Publisher).
TONIN, M. and VLASSOPOULOS, M. (2015). "Corporate Philanthropy and Productivity: Evidence from an Online Real Effort Experiment", *Management Science*, **61**, 1795–1811.
TVERSKY, A. and KAHNEMAN, D. (1991), "Advances in Prospect Theory: Cumulative Representation of Uncertainty", *Journal of Risk and Uncertainty*, **5**, 297–323.
VIVALT, E. (2016), "How Much Can We Generalize from Impact Evaluations?" (Working Paper).
WU, G. and GONZALEZ, R. (1996), "Curvature of the Probability Weighting Function", *Management Science*, **42**, 1676–1690.